3

An Overview of the Bayesian Approach

In this chapter we shall introduce the core issues of Bayesian reasoning: these include subjectivity and context, the use of Bayes theorem, Bayes factors, interpretation of study results, prior distributions, predictions, decision-making, multiplicity, using historical data, and computation. This overview necessarily covers a wide range of material and ideas at an introductory level, and the issues will be further developed in subsequent chapters. A structure for reporting Bayesian analyses is proposed, which will provide a uniform style for the examples presented in this book. A number of starred sections can be omitted without loss of continuity.

3.1 SUBJECTIVITY AND CONTEXT

The standard interpretation of probability describes long-run properties of repeated random events (Section 2.1.1). This is known as the *frequency* interpretation of probability, and standard statistical methods are sometimes referred to as 'frequentist'. In contrast, the Bayesian approach rests on an essentially 'subjective' interpretation of probability, which is allowed to express generic uncertainty or 'degree of belief' about any unknown but potentially observable quantity, whether or not it is one of a number of repeatable experiments. For example, it is quite reasonable from a subjective perspective to think of a probability of the event 'Earth will be openly visited by aliens in the next ten years', whereas it may be difficult to interpret this potential event as part of a 'long-run' series. Methods of assessing subjective probabilities and probability distributions will be discussed in Section 5.2.

The rules of probability listed in Section 2.1.1 are generally taken as self-evident, based on comparison with simple chance situations such as rolling dice or drawing coloured balls out of urns. In these experiments there will be a

Bayesian Approaches to Clinical Trials and Health-Care Evaluation D. J. Spiegelhalter, K. R. Abrams and J. P. Myles
© 2004 John Wiley & Sons, Ltd ISBN: 0-471-49975-7

general consensus about the probabilities due to assumptions about physical symmetries: if a balanced coin is to be tossed, the probability of it coming up 'heads' will usually be assigned 0.5, whether this is taken as a subjective belief about the next toss or whether the next toss is thought of as part of a long series of tosses. However, as Lindley (2000) emphasises, the rules of probability do not need to be assumed as self-evident, but can be *derived* from 'deeper' axioms of reasonable behaviour of an individual (say, *You*) in the face of Your own uncertainty. This 'reasonable behaviour' features characteristics such as Your unwillingness to make a series of bets based on expressed probabilities, such that You are bound to lose (a so-called 'Dutch book'), or Your unwillingness to state probabilities that can always be improved upon in terms of their expected accuracy in predicting events. It is perhaps remarkable that from such conditions one can prove the three basic rules of probability (Lindley, 1985): as a simple example, if I state probabilities of 0.7 that it will rain tomorrow, and 0.4 that it will not rain, and I am willing to bet at these odds, then a good bookmaker can accept a series of bets from me such that I am bound to lose. (For example, assuming small stakes, I would consider it a good deal to bet 14 units of money for a return of 21 if it rained, since my expected profit is $0.7 \times 21 - 14 = 0.7$, and simultaneously I would bet 8 units of money for a return of 21 if it did *not* rain. Thus the bookmaker is certain to make a profit of 1 unit whatever happens.) Such probabilities are said not to 'cohere', and are assumed to be avoided by all rational individuals.

The vital point of the subjective interpretation is that Your probability for an event is a property of Your relationship to that event, and not an objective property of the event itself. This is why, pedantically speaking, one should always refer to probabilities *for* events rather than probabilities *of* events, and the conditioning context H used in Section 2.1.1 includes the observer and all their background knowledge and assumptions. The fact that the probability is a reflection of personal uncertainty rather than necessarily being based on future unknown events is illustrated (from personal experience) by a gambling game played in casinos in Macau. Two dice are thrown out of sight of the gamblers and immediately covered up: the participants then bet on different possible combinations. Thus, they are betting on an event that has already occurred, but about which they are personally ignorant. (Incidentally, their beliefs also do not appear to be governed by the assumed physical symmetries of the dice: although they have 2 minutes to bet, everyone remains totally still for at least 90 seconds, and then when the first bet is laid the crowd follow in a rush, apparently believing in the good fortune of the one confident individual.)

The subjective view of probability is not new, and in past epochs has been the standard ideology. Fienberg (1992) points out that Jakob Bernoulli in 1713 introduced 'the subjective notion that the probability is personal and varies with an individual's knowledge', and that Laplace and Gauss both worked with posterior distributions two hundred years ago, which became known as 'the inverse method'. However, from the mid-nineteenth century the frequency approach started to dominate, and controversy has sporadically continued.

Dempster (1998) quotes Edgeworth in 1884 as saying that the critics who 'heaped ridicule upon Bayes' theorem and the inverse method' were trying to elicit 'knowledge out of ignorance, something out of nothing'. Polemical opinions are still expressed in defence of the explicit introduction of subjective judgement into scientific research: 'it simply makes no sense to take seriously every apparent falsification of a plausible theory, any more than it makes sense to take seriously every new scientific idea' (Matthews, 1998).

Bayesian methods therefore explicitly allow for the possibility that the conclusions of an analysis may depend on who is conducting it and their available evidence and opinion, and therefore the context of the study is vital: 'Bayesian statistics treats subjectivity with respect by placing it in the open and under the control of the consumer of data' (Berger and Berry, 1988). Apart from methodological researchers, at least five different viewpoints might be identified for an evaluation of a health-care intervention:

- *sponsors*, e.g. the pharmaceutical industry, medical charities or granting agencies;
- *investigators*, *i.e.* those responsible for the conduct of a study, whether industry or publicly funded;
- *reviewers*, e.g. regulatory bodies;
- *policy makers*, e.g. agencies setting health policy;
- *consumers*, e.g. individual patients or clinicians acting on their behalf.

Each of these broad categories can be further subdivided. An analysis which might be carried out solely for the investigators, for example, may not be appropriate for presentation to reviewers or consumers: 'experimentalists tend to draw a sharp distinction between providing their opinions and assessments for the purposes of experimental design and in-house discussion, and having them incorporated into any form of externally disseminated report' (Racine *et al*., 1996). The roles of these different stakeholders in decision-making is further explored in Chapter 9.

A characteristic of health-care evaluation is that the investigators who plan and conduct a study are generally not the same body as those who make decisions on the basis of the evidence provided in part by that study: such decision-makers may be regulatory authorities, policy-makers or health-care providers. This division is acknowledged in this book by separating Chapter 6 on the design and monitoring of trials from Chapter 9 on policy-making.

3.2 BAYES THEOREM FOR TWO HYPOTHESES

In Section 2.1.3 Bayes theorem was derived as a basic result in probability theory. We now begin to illustrate its use as a mechanism for learning about unknown quantities from data, a process which is sometimes known as 'prior to posterior' analysis. We start with the simplest possible situation. Consider two hypotheses H_0 and H_1 which are 'mutually exhaustive and exclusive', *i.e.* one and only one is true. Let the *prior* probability for each of the two hypotheses, before we have access to the evidence of interest, be $p(H_0)$ and $p(H_1)$; for the moment we will not concern ourselves with the source of those probabilities. Suppose we have observed some data y, such as the results of a test, and we know from past experience that the probability of observing y under each of the two hypotheses is $p(y|H_0)$ and $p(y|H_1)$, respectively: these are the *likelihoods*, with the vertical bar representing 'conditioning'.

Bayes theorem shows how to revise our prior probabilities in the light of the evidence in order to produce *posterior probabilities*. Specifically, by adapting (2.3) we have the identity

$$p(H_0|y) = \frac{p(y|H_0)}{p(y)} \times p(H_0), \tag{3.1}$$

where $p(y) = p(y|H_0)p(H_0) + p(y|H_1)p(H_1)$ is the overall probability of y occurring.

Now $H_1 =$ 'not H_0' and so $p(H_0) = 1 - p(H_1)$ and $p(H_0|y) = 1 - p(H_1|y)$. In terms of odds rather than probabilities, Bayes theorem can then be re-expressed (see (2.5)) as

$$\frac{p(H_0|y)}{p(H_1|y)} = \frac{p(y|H_0)}{p(y|H_1)} \times \frac{p(H_0)}{p(H_1)}. \tag{3.2}$$

Now $p(H_0)/p(H_1)$ is the 'prior odds', $p(H_0|y)/p(H_1|y)$ is the 'posterior odds', and $p(y|H_0)/p(y|H_1)$ is the ratio of the likelihoods, and so (3.2) can be expressed as

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}.$$

By taking logarithms we also note that

$$\log(\text{posterior odds}) = \log(\text{likelihood ratio}) + \log(\text{prior odds}).$$

where the log(likelihood ratio) has also been termed the 'weight of evidence': this term was invented by Alan Turing when using these techniques for breaking the Enigma codes at Bletchley Park during the Second World War.

Example 3.1 shows how this formulation is commonly used in the evaluation of diagnostic tests, and reveals that our intuition is often poor when processing probabilistic evidence, and that we tend to forget the importance of the prior probability (Section 5.2).

Example 3.1 *Diagnosis: Bayes theorem in diagnostic testing*

Suppose a new home HIV test is claimed to have '95% sensitivity and 98% specificity', and is to be used in a population with an HIV prevalence of

1/1000. We can calculate the expected status of 100 000 individuals who are tested, and the results are shown in Table 3.1. Thus, for example, we expect 100 truly HIV positive individuals of whom 95% will test positive, and of the remaining 99 900 HIV negative individuals we expect 2% (1998) to test positive. Thus of the 2093 who test positive (*i.e.* have observation y), only 95 are truly HIV positive, giving a 'predictive value positive' of only $95/2093 = 4.5\%$.

Table 3.1 Expected status of 100 000 tested individuals in a population with an HIV prevalence of 1/1000.

	HIV−	HIV+	
Test −	97 902	5	97 907
Test +	1 998	95	2 093
	99 900	100	100 000

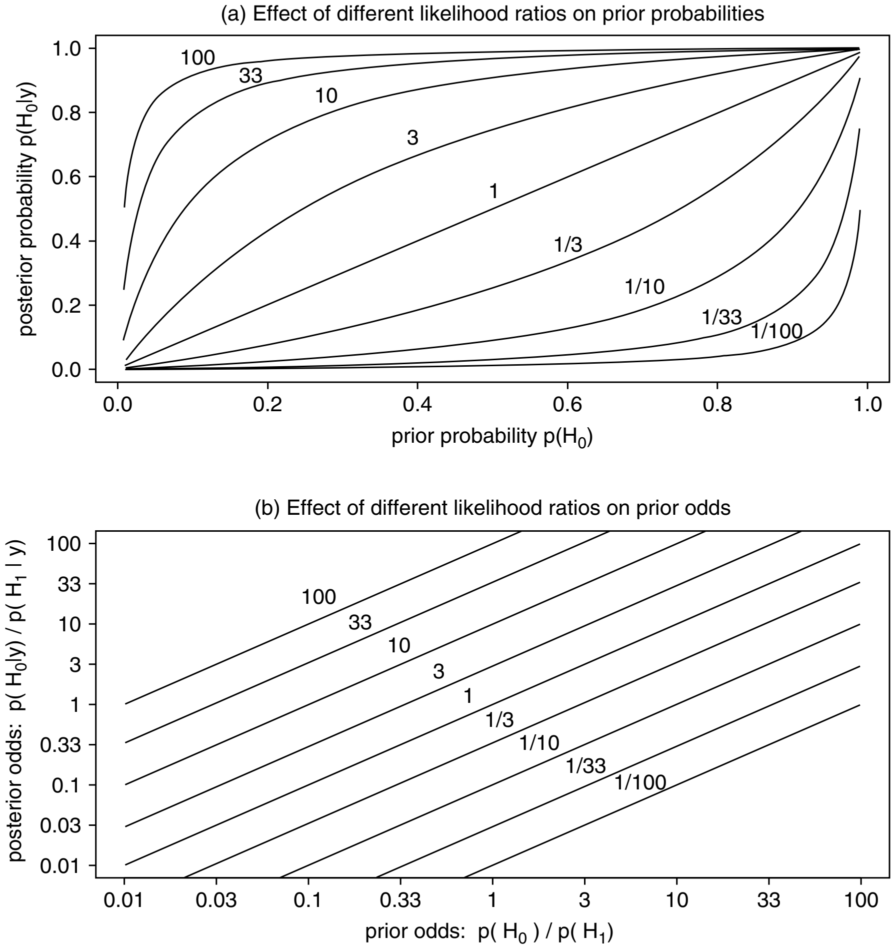
We can also do these calculations using Bayes theorem. Let H_0 be the hypothesis that the individual is truly HIV positive, and y be the observation that they test positive. The disease prevalence is the prior probability ($p(H_0) = 0.001$), and we are interested in the chance that someone who tests positive is truly HIV positive, *i.e.* the posterior probability $p(H_0|y)$.

Let H_1 be the hypothesis that they are truly HIV negative; '95% sensitivity' means that $p(y|H_0) = 0.95$, and '98% specificity' means that $p(y|H_1) = 0.02$. To use (3.2), we require two inputs: the prior odds $p(H_0)/p(H_1)$ which are 1/999, and the likelihood ratio $p(y|H_0)/p(y|H_1)$ which is $0.95/0.02 = 95/2$. Then from (3.2) the posterior odds are $(95/2) \times 1/999 = 95/1998$. These odds correspond to a posterior probability $p(H_0|y) = 95/(95 + 1998) = 0.045$, as found directly from the table.

Alternatively, we can use the form of Bayes theorem given by (3.1). Now $p(y) = p(y|H_0)p(H_0) + p(y|H_1)p(H_1) = 0.95 \times 0.001 + 0.02 \times 0.999 = 0.020\,93$. Thus (3.1) says that $p(H_0|y) = 0.95 \times 0.001/0.020\,93 = 0.045$.

The crucial finding is that over 95% of those testing positive will, in fact, not have HIV.

Figure 3.1 shows Bayes theorem for two hypotheses in either odds or probability form, for a range of likelihood ratios. The likelihood ratio from a positive result in Example 3.1 is $0.95/0.02 = 47.5$. From a rough inspection of Figure 3.1 we can see that such a likelihood ratio is sufficient to turn a moderately low prior probability, such as 0.2, into a reasonably high posterior probability of around 0.9; however, if the prior probability is as low as it is in Example 3.1 (*i.e.* 0.001), then the posterior probability is still somewhat small.



(a) Effect of different likelihood ratios on prior probabilities

(b) Effect of different likelihood ratios on prior odds

Figure 3.1 Bayes theorem for two hypotheses H_0 and H_1 = 'not H_0' in (a) probability $p(H_0)$ and (b) odds $p(H_0)/p(H_1)$ form. By specifying the prior probability or odds, and the likelihood ratio $p(y|H_0)/p(y|H_1)$, the posterior probability or odds can be read off the graph. Note that (b) uses a logarithmic scaling, under which Bayes theorem gives a linear relationship.

3.3 COMPARING SIMPLE HYPOTHESES: LIKELIHOOD RATIOS AND BAYES FACTORS

In Section 3.2 we showed how data y influence the relative probabilities of two hypotheses H_0 and H_1 through the likelihood ratio $p(y|H_0)/p(y|H_1)$, and hence the likelihoods contain all the relevant evidence that can be extracted from the data: this is the likelihood principle, discussed in more detail in Section 4.3. This measure of the relative likelihood of two hypotheses is also known as the 'Bayes factor' (BF), although Cornfield (1976) also termed this the 'relative betting odds' between two hypotheses: see, for example, Goodman (1999b) for a detailed exposition. The Bayes factor can vary between 0 and ∞, with small values being considered as both evidence *against* H_0 and evidence *for* H_1. The scale in Table 3.2 was provided by the Bayesian physicist, Harold Jeffreys, and dates from 1939 (Jeffreys, 1961, p. 432).

The crucial idea is that the Bayes factor transforms prior to posterior odds: this uses expression (3.2), and the results can be read off Figure 3.1. In Example 3.1 we observed a Bayes factor (likelihood ratio) after a positive HIV test of $\mathrm{BF} = 47.5$ in favour of being HIV positive (H_0). Table 3.2 labels this as 'very strong' evidence in itself in favour of H_0, but when combined with strong prior opinion against H_0 (prior odds of 1/999) does not lead to a very convincing result (posterior odds $\approx 1/21$).

Bayes factors can also be obtained for composite hypotheses that include unknown parameters: this is discussed in Section 4.4 and is a feature when using a prior distribution that puts a 'lump' of probability on a (null) hypothesis (Section 5.5.4). The relationship between Bayes factors and traditional ways of hypothesis testing has been the subject of considerable research and controversy, and is discussed further in Section 4.4.

The use of Bayes theorem in diagnostic testing is an established part of formal clinical reasoning. More controversial is the use of Bayes theorem in general statistical analyses, where a parameter θ is an unknown quantity such as the mean benefit of a treatment on a specified patient population, and its prior distribution $p(\theta)$ needs to be specified. This major step might be considered as a natural extension of the subjective interpretation of probability, but the following (starred) section provides a further argument for why a prior distribution on a parameter may be a reasonable assumption.

Table 3.2 Calibration of Bayes factor (likelihood ratio) provided by Jeffreys.

Bayes factor range	Strength of evidence in favour of H_0 and against H_1
> 100	Decisive
32 to 100	Very strong
10 to 32	Strong
3.2 to 10	Substantial
1 to 3.2	'Not worth more than a bare mention'
	Strength of evidence against H_0 and in favour of H_1
1 to 1/3.2	'Not worth more than a bare mention'
1/3.2 to 1/10	Substantial
1/10 to 1/32	Strong
1/32 to 1/100	Very strong
$< 1/100$	Decisive

3.4 EXCHANGEABILITY AND PARAMETRIC MODELLING*

In Section 2.2.3 we introduced the concept of independent and identically distributed (i.i.d.) variables $Y_1, \ldots, Y_n$ as a fundamental component of standard statistical modelling. However, just as we found in Section 3.1 that the rules of probability could themselves be derived from more basic ideas of rational behaviour, so we can derive the idea of i.i.d. variables and prior distributions of parameters from the more basic subjective judgement known as 'exchangeability'. Exchangeability is a formal expression of the idea that we find no systematic reason to distinguish the individual variables $Y_1, \ldots, Y_n$ – they are similar but not identical. Technically, we judge that $Y_1, \ldots, Y_n$ are exchangeable if the probability that we assign to any set of potential outcomes, $p(y_1, \ldots, y_n)$, is unaffected by permutations of the labels attached to the variables. For example, suppose Y_1, Y_2, Y_3 are the first three tosses of a (possibly biased) coin, where $Y_1 = 1$ indicates a head, and $Y_1 = 0$ indicates a tail. Then we would judge $p(Y_1 = 1,\ Y_2 = 0,\ Y_3 = 1) = p(Y_2 = 1,\ Y_1 = 0,\ Y_3 = 1) = p(Y_1 = 1, Y_3 = 0, Y_2 = 1)$, *i.e.* the probability of getting two heads and a tail is unaffected by the particular toss on which the tail comes. This is a natural judgement to make if we have no reason to think that one toss is systematically any different from another. Note that it does *not* mean we believe that $Y_1, \ldots, Y_n$ are independent: independence would imply $p(y_1, \ldots, y_n) = p(y_1) \times \ldots \times p(y_n)$ and hence the result of a series of tosses does not help us predict the next, whereas a long series of heads would tend to make us believe the coin was seriously biased and hence would lead us to predict a head as more likely.

An Italian actuary, Bruno de Finetti, published in 1930 a most extraordinary result (de Finetti, 1930). He showed that if a set of binary variables $Y_1, \ldots, Y_n$ were judged exchangeable, then it implied that

$$p(y_1, \ldots, y_n) = \int \prod_{i=1}^{n} p(y_i|\theta)p(\theta)d\theta. \tag{3.3}$$

Now (3.3) is unremarkable if we argue from right to left: if $Y_1, \ldots, Y_n$ are i.i.d., each with distribution $p(y_i|\theta)$, their joint distribution (conditional on θ) is $p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$ (2.16). Hence, their marginal distribution $p(y_1, \ldots, y_n)$ (2.7), given a distribution $p(\theta)$, is given by (3.3). However, de Finetti's remarkable achievement was to argue from left to right: exchangeable random quantities can be thought of as being i.i.d. variables drawn from some common distribution depending on an unknown parameter θ, which itself has a prior distribution $p(\theta)$. Thus, from a subjective judgement about observable quantities, one derives the whole apparatus of i.i.d. variables, conditional independence, parameters and prior distributions. This was an amazing achievement.

De Finetti's results have been extended to much more general situations (Bernardo and Smith, 1994), and the concept of exchangeability will continually recur throughout this book.

3.5 BAYES THEOREM FOR GENERAL QUANTITIES

This small section is the most important in this book.

Suppose θ is some quantity that is currently unknown, for example the true success rate of a new therapy, and let $p(\theta)$ denote the prior distribution of θ. As discussed in Section 3.1, this prior distribution should, strictly speaking, be denoted $p(\theta|H)$ to remind us that it represents Your judgement about θ conditional on a context H, where You are the person for whom the analysis is being performed (the client), and not the statistician who may be actually carrying out the analysis. The interpretation and source of such distributions are discussed in Section 3.9 and Chapter 5.

Suppose we have some observed evidence y, for example the results of a clinical trial, whose probability of occurrence is assumed to depend on θ. As we have seen, this dependence is formalised by $p(y|\theta)$, the (conditional) probability of y for each possible value of θ, and when considered as a function of θ is known as the likelihood. We would like to obtain the new, posterior, probability for different values of θ, taking account of the evidence y; this probability has the conditioning reversed and is denoted $p(\theta|y)$.

Bayes theorem applied to a general quantity θ was given in (2.6) and says that

$$p(\theta|y) = \frac{p(y|\theta)}{p(y)} \times p(\theta). \tag{3.4}$$

Now $p(y)$ is just a normalising factor to ensure that $\int p(\theta|y)\, d\theta = 1$, and its value is not of interest (unless we are comparing alternative models). The essence of Bayes theorem only concerns the terms involving θ, and hence it is often written

$$p(\theta|y) \propto p(y|\theta) \times p(\theta), \tag{3.5}$$

which says that the posterior distribution is proportional to (*i.e.* has the same shape as) the product of the likelihood and the prior. The deceptively simple expression (3.5) is the basis for the whole of the rest of this book, since it shows how to make inferences from a Bayesian perspective, both in terms of estimation and obtaining credible intervals and also making direct probability statements about the quantities in which we are interested.

3.6 BAYESIAN ANALYSIS WITH BINARY DATA

In Section 2.2.4 we considered a probability θ of an event occurring, and derived the form of the likelihood for θ having observed n cases in which r events occurred. Adopting a Bayesian approach to making inferences, we wish to combine this likelihood with initial evidence or opinion regarding θ, as expressed in a prior distribution $p(\theta)$.

3.6.1 Binary data with a discrete prior distribution

First, suppose only a limited set of hypotheses concerning the true proportion θ are being entertained, corresponding to a finite list denoted $\theta_1, \ldots, \theta_J$. Suppose in addition a prior probability $p(\theta_j)$ of each has been assessed, where $\sum_j p(\theta_j) = 1$. For a single Bernoulli trial with outcome 0 or 1, the likelihood for each possible value for θ is given by (2.15),

$$p(y|\theta_j) = \theta_j^y (1-\theta_j)^{1-y}, \tag{3.6}$$

i.e. $p(y|\theta_j) = \theta_j$ if $y = 1$, and $p(y|\theta_j) = 1 - \theta_j$ if $y = 0$.

Having observed an outcome y, Bayes theorem (3.5) states that the posterior probabilities for the θ_j obey

$$p(\theta_j|y) \propto \theta_j^y (1-\theta_j)^{1-y} \times p(\theta_j), \tag{3.7}$$

where the normalising factor that ensures that the posterior probabilities add to 1 is

$$p(y) = \textstyle\sum_j \theta_j^y (1-\theta_j)^{1-y} \times p(\theta_j).$$

After further observations have been made, say with the result that there have been r ‘successes’ out of n trials, the relevant posterior will obey

$$p(\theta_j|r) \propto \theta_j^r (1-\theta_j)^{n-r} \times p(\theta_j). \tag{3.8}$$

A basic example of these calculations is given in Example 3.2.

Example 3.2 *Drug: Binary data and a discrete prior*

Suppose a drug has an unknown true response rate θ, and for simplicity we assume that θ can only take one of the values $\theta_1 = 0.2$, $\theta_2 = 0.4$, $\theta_3 = 0.6$ or $\theta_4 = 0.8$. Before experimentation we adopt the ‘neutral’ position of assuming each value θ_j is equally likely, so that $p(\theta_j) = 0.25$ for each $j = 1, 2, 3, 4$.

Suppose we test the drug on a single subject and we observed a positive response ($y = 1$). How should our belief in the possible values of θ be revised?

First, we note that the likelihood is simply $p(y|\theta_j) = \theta_j^y (1-\theta)^{(1-y)} = \theta_j$. Table 3.3 displays the components of Bayes theorem (3.7): the ‘Likelihood × prior’ column, normalised by its sum $p(y)$, gives the posterior probabilities. It is perhaps initially surprising that a single positive response makes it four times as likely that the true response rate is 80% rather than 20%.

Table 3.3 Results after observing a single positive response, $y = 1$, for a drug given an initial uniform distribution over four possible response rates θ_j.

j	θ_j	Prior $p(\theta_j)$	Likelihood $p(y\|\theta_j)$	Likelihood × prior $p(y\|\theta_j)p(\theta_j)$	Posterior $p(\theta_j\|y)$
1	0.2	0.25	0.2	0.05	0.10
2	0.4	0.25	0.4	0.10	0.20
3	0.6	0.25	0.6	0.15	0.30
4	0.8	0.25	0.8	0.20	0.40
	$\sum_j$	1.0		0.50	1.0

Suppose we now observe 15 positive responses out of 20 patients, how is our belief revised? Table 3.4 shows that any initial belief in $\theta_1 = 0.2$ is now completely overwhelmed by the data, and that the only remaining contenders are $\theta_3 = 0.6$ with about 30% of the posterior probability, and $\theta_4 = 0.8$ with about 70%.

We note that, had we given any non-zero probability to the extreme values of $\theta = 0,\ 1$, *i.e.* the drug either never or always worked, these would give a zero likelihood and hence zero posterior probability.

Table 3.4 Results after observing 15 positive responses, $y = 15$, for a drug out of 20 cases, given an initial uniform distribution over four possible response rates θ_j.

j	θ_j	Prior $p(\theta_j)$	Likelihood $\theta_j^{15}(1-\theta_j)^5$ $(\times 10^{-7})$	Likelihood × prior $\theta_j^{15}(1-\theta_j)^5\ p(\theta_j)$ $(\times 10^{-7})$	Posterior $p(\theta_j\|X = 1)$
1	0.2	0.25	0.0	0.0	0.000
2	0.4	0.25	0.8	0.2	0.005
3	0.6	0.25	48.1	12.0	0.298
4	0.8	0.25	112.6	28.1	0.697
	$\sum_j$	1.0		40.3	1.0

3.6.2 Conjugate analysis for binary data

It is generally more realistic to consider θ a continuous parameter, and hence it needs to be given a continuous prior distribution. One possibility is that we think all possible values of θ are equally likely, in which case we could summarise this by a uniform distribution (Section 2.6.4) so that $p(\theta) = 1$ for $0 \leqslant \theta \leqslant 1$. Applying Bayes theorem (3.5) yields

$$p(\theta|y) \propto \theta^r(1-\theta)^{n-r} \times 1, \tag{3.9}$$

where r is the number of events observed and n is the total number of individuals.

We may recognise that the functional form of the posterior distribution in (3.9) is proportional to that of a beta distribution (Section 2.6.3). Rewriting the posterior distribution (3.9) as $\theta^{(r+1)-1}(1-\theta)^{(n-r+1)-1}$, we can see that the posterior distribution is in fact Beta $[r+1,\ n-r+1]$. This immediately means that we can now summarise the posterior distribution in terms of its mean and variance, and make probability statements based on what we know about the beta distribution (for example, many common statistical packages will calculate tail area probabilities for the beta distribution).

Instead of a uniform prior distribution for θ we could take a Beta $[a, b]$ prior distribution and obtain the following analysis:

$$\begin{aligned}
\text{Prior} &\propto \theta^{a-1}(1-\theta)^{b-1} \\
\text{Likelihood} &\propto \theta^r(1-\theta)^{n-r} \\
\text{Posterior} &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^r(1-\theta)^{n-r} \\
&\propto \theta^{a+r-1}(1-\theta)^{b+n-r-1} \\
&= \text{Beta}[a+r,\ b+n-r].
\end{aligned} \tag{3.10}$$

Thus we have specified a beta prior distribution for a parameter, observed data from a Bernoulli or binomial sampling distribution, worked through Bayes theorem, and ended up with a beta posterior distribution. This is a case of *conjugate analysis*. Conjugate models occur when the posterior distribution is of the same *family* as the prior distribution: other examples include the gamma distribution being conjugate with a Poisson likelihood, normal priors being conjugate with normal likelihoods (Section 3.7), and gamma priors for unknown precisions of normal likelihoods (Section 2.6.5).

Example 3.3 *Drug (continued): Binary data and a continuous prior*

Suppose that previous experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible, with an expectation around 0.4. We can translate this into a prior Beta$[a, b]$ distribution as follows.

We first want to estimate the mean m and standard deviation s of the prior distribution. For normal distributions we know that $m \pm 2s$ includes just over 95% of the probability, so if we were assuming a normal prior we might estimate $m = 0.4,\ s = 0.1$. However, we know from Section 2.6.3 that beta distributions with reasonably high a and b have an approximately normal shape, so these estimates might also be used for a beta prior.

Next, from Section 2.6.3, we know that for a beta distribution

$$m = a/(a+b), \tag{3.11}$$

$$s^2 = m(1-m)/(a+b+1). \tag{3.12}$$

Expression (3.12) can be rearranged to give $a + b = m(1 - m)/s^2 - 1$. Using the estimates $m = 0.4$, $s = 0.1$, we obtain $a + b = 23$. Then, from (3.11), we see that $a = m(a + b)$, and hence we finally obtain $a = 9.2$, $b = 13.8$: this can be considered a 'method of moments'. A Beta[9.2,13.8] distribution is shown in Figure 3.2(a), showing that it well represents the prior assumptions. It is convenient to think of this prior distribution as that which would have arisen had we started with a 'non-informative' prior Beta[0,0] and then observed $a = 9.2$ successes in $a + b = 23$ patients (however, this is only a heuristic argument as there is no agreed 'non-informative' beta prior, with Beta[0,0], Beta[$\frac{1}{2}$,$\frac{1}{2}$], Beta[1,1] all having been suggested (Section 5.5.1)).

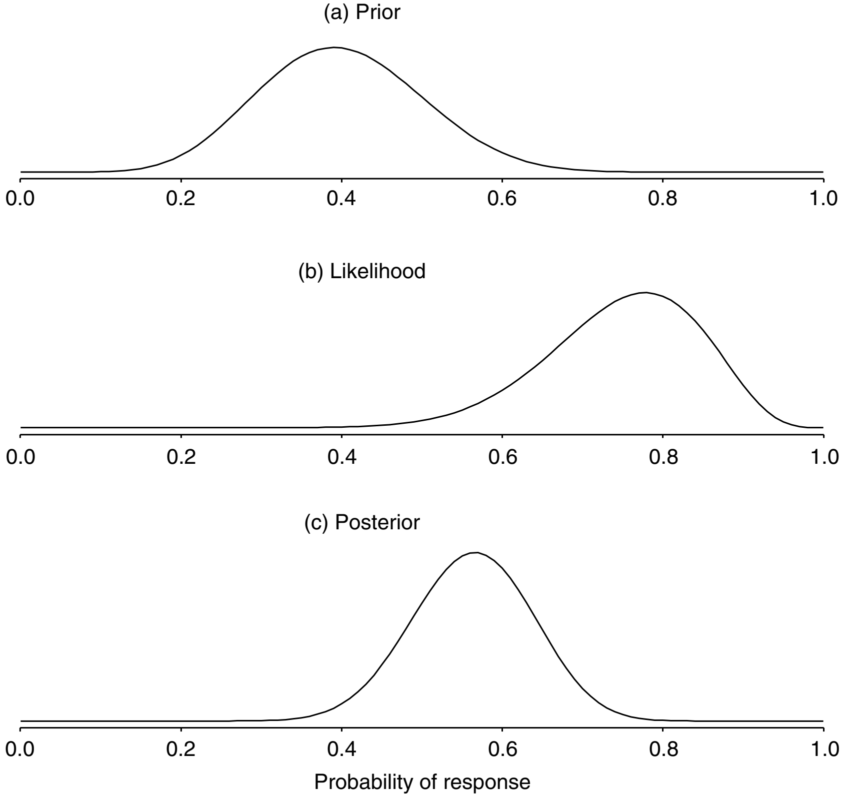


Figure 3.2 (a) is a Beta[9.2,13.8] prior distribution supporting response rates between 0.2 and 0.6, (b) is a likelihood arising from a binomial observation of 15 successes out of 20 cases, and (c) is the resulting Beta[24.2, 18.8] posterior from a conjugate beta-binomial analysis.

If we now observed $r = 15$ successes out of 20 trials, we know from (3.10) that the parameters of the beta distribution are updated to $[a + 15,\ b + 20 - 5] = [24.2,\ 18.8]$. The likelihood and posterior are shown in Figures 3.2(b) and 3.2(c): the posterior will have mean $24.2/(24.2 + 18.8) = 0.56$.

3.7 BAYESIAN ANALYSIS WITH NORMAL DISTRIBUTIONS

In Section 2.4 we saw that in many circumstances it is appropriate to consider a likelihood as having a normal shape, although this may involve working on somewhat uninituitive scales such as the logarithm of the hazard ratio. With a normal likelihood it is mathematically convenient, and often reasonably realistic, to make the assumption that the prior distribution $p(\theta)$ has the form

$$p(\theta) = \mathrm{N}\left[\theta \middle| \mu,\ \frac{\sigma^2}{n_0}\right], \tag{3.13}$$

where μ is the prior mean. We note that the same standard deviation σ is used in the likelihood and the prior, but the prior is based on an 'implicit' sample size n_0. The advantage of this formulation becomes apparent when we carry out prior-to-posterior analysis. We note in passing that as n_0 tends to 0, the variance becomes larger and the distribution becomes 'flatter', and in the limit the distribution becomes essentially uniform over $(-\infty,\ \infty)$. A normal prior with a very large variance is sometimes used to represent a 'non-informative' distribution (Section 5.5.1).

Suppose we assume such a normal prior $\theta \sim \mathrm{N}[\mu,\ \sigma^2/n_0]$ and likelihood $y_m \sim \mathrm{N}[\theta,\ \sigma^2/m]$. Then the posterior distribution obeys

$$\begin{aligned} p(\theta|y_m) &\propto p(y_m|\theta)p(\theta) \\ &\propto \exp\left[-\frac{(y_m - \theta)^2 m}{2\sigma^2}\right] \times \exp\left[-\frac{(\theta - \mu)^2 n_0}{2\sigma^2}\right], \end{aligned}$$

ignoring irrelevant terms that do not include θ. By matching terms in θ it can be shown that

$$(y_m - \theta)^2 m + (\theta - \theta_0)^2 n_0 = \left(\theta - \frac{n_0\theta_0 + m y_m}{n_0 + m}\right)^2 (n_0 + m) + (y_m - \mu)^2\left(\frac{1}{m} + \frac{1}{n_0}\right),$$

and we can recognise that the term involving θ is exactly that arising from a posterior distribution

$$p(\theta|y_m) = \mathrm{N}\left[\theta \middle| \frac{n_0\mu + my_m}{n_0+m}, \frac{\sigma^2}{n_0+m}\right]. \tag{3.14}$$

Equation (3.14) is very important. It says that our posterior mean $(n_0\mu + my_m)/(n_0+m)$ is a weighted average of the prior mean μ and parameter estimate y_m, weighted by their precisions, and therefore is always a compromise between the two. Our posterior variance (1/precision) is based on an implicit sample size equivalent to the sum of the prior 'sample size' n_0 and the sample size of the data m: thus, when combining sources of evidence from the prior and the likelihood, we *add precisions* and hence always decrease our uncertainty. As Senn (1997a, p. 46) claims, 'A Bayesian is one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule'. Note that as $n_0 \to 0$, the prior tends towards a uniform distribution and the posterior tends to the same shape as the likelihood.

Suppose we do not adopt the convention for expressing prior and sampling variances as σ^2/n_0 and σ^2/m, and instead use the general notation $\theta \sim \mathrm{N}[\mu, \tau^2]$ and likelihood $y_m \sim \mathrm{N}[\theta, \sigma_m^2]$. Then it is straightforward to show that the posterior distribution (3.14) can be expressed as

$$p(\theta|y_m) = \mathrm{N}\left[\theta \middle| \frac{\frac{\mu}{\tau^2} + \frac{y_m}{\sigma_m^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma_m^2}}, \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma_m^2}}\right]. \tag{3.15}$$

We will sometimes find this general form useful, but will generally find (3.14) more intuitive.

Example 3.4 provides a simple example of Bayesian reasoning using normal distributions.

Example 3.4 *SBP: Bayesian analysis for normal data*

Suppose we are interested in the long-term systolic blood pressure (SBP) in mmHg of a particular 60-year-old female. We take two independent readings 6 weeks apart, and their mean is 130. We know that SBP is measured with a standard deviation $\sigma = 5$. What should we estimate her SBP to be?

Let her long-term SBP be denoted θ. A standard analysis would use the sample mean $y_m = 130$ as an estimate, with standard error $\sigma/\sqrt{m} = 5/\sqrt{2} = 3.5$: a 95% confidence interval is $y_m \pm 1.96 \times \sigma/\sqrt{m}$, *i.e.* 123.1 to 136.9.

However, we may have considerable additional information about SBPs which we can express as a prior distribution. Suppose that a survey in the same population revealed that females aged 60 had a mean long-term SBP of 120 with standard deviation 10. This population distribution can be considered as a prior distribution for the specific individual, and is shown in Figure 3.3(a): if we express the prior standard deviation as $\sigma/\sqrt{n_0}$ (*i.e.* variance σ^2/n_0), we can solve to find $n_0 = (\sigma/10)^2 = 0.25$.

Figure 3.3(b) shows the likelihood arising from the two observations on the woman. From (3.14) the posterior distribution of θ is normal with mean $(0.25 \times 120 + 2 \times 130)/(0.25 + 2) = 128.9$ and standard deviation $\sigma/\sqrt{n_0 + m} = 5/\sqrt{2.25} = 3.3$, giving a 95% interval of $128.9 \pm 1.96 \times 3.3 = (122.4,\ 135.4)$. Figure 3.3(c) displays this posterior distribution, revealing some 'shrinkage' towards the population mean, and a small increase in precision from not using the data alone.

Intuitively, we can say that the woman has somewhat higher measurements than we would expect for someone her age, and hence we slightly adjust our estimate to allow for the possibility that her two measures happened by chance to be on the high side. As additional measures are made, this possibility becomes less plausible and the prior knowledge will be systematically downgraded.

3.8 POINT ESTIMATION, INTERVAL ESTIMATION AND INTERVAL HYPOTHESES

Although it is most informative to plot an entire posterior distribution, there will generally be a need to produce summary statistics: we shall consider point estimates, intervals, and the probabilities of specified hypotheses.

Point estimates. Traditional measures of location of distributions include the mean, median and mode, and – by imposing a particular penalty on error in estimation (Berger, 1985) – each can be given a theoretical justification as a point estimate derived from a posterior distribution. If the posterior distribution is symmetric and unimodal, as in Figure 3.3, then the mean, median and mode all coincide in a single value and there is no difficulty in making a choice. We shall find, however, that in some circumstances posterior distributions are considerably skewed and there are marked differences between, say, mean and median. We shall prefer to quote the median in such contexts as it is less sensitive to the tails of the distribution, although it is perhaps preferable to report all three summary measures when they show wide disparity.

Interval estimates. Any interval containing, say, 95% probability may be termed a 'credible' interval to distinguish it from a Neyman–Pearson 'confidence interval', although we shall generally refer to them simply as posterior intervals. Three types of intervals can be distinguished – we assume a continuous parameter θ with range on $(-\infty,\ \infty)$ and a posterior conditional on generic data y:

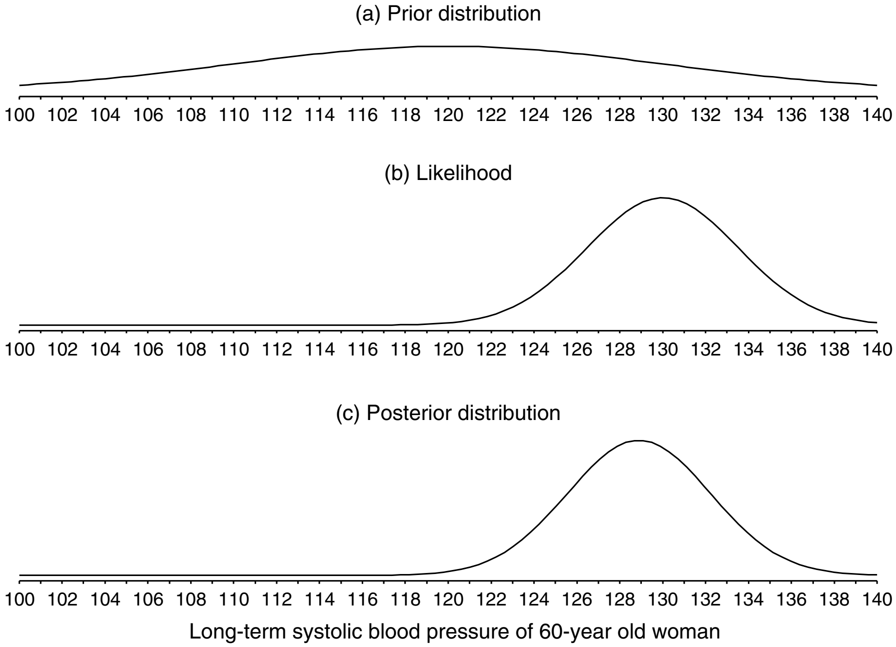


Figure 3.3 Estimating the true long-term underlying systolic blood pressure of a 60-year-old woman: (a) the prior distribution is N[120, 10^2] and expresses the distribution of true SBPs in the population; (b) the likelihood is proportional to N[130, 3.5^2] and expresses the support for different values arising from the two measurements made on the woman; (c) the posterior distribution is N[128.9, 3.3^2] and is proportional to the likelihood multiplied by the prior.

One-sided intervals. For example, a one-sided upper 95% interval would be (θ_L, ∞), where $p(\theta < \theta_L|y) = 0.05$.

Two-sided 'equi-tail-area' intervals. A two-sided 95% interval with equal probability in each tail area would comprise (θ_L, θ_U), where $p(\theta < \theta_L|y) = 0.025$, and $p(\theta > \theta_U|y) = 0.975$.

Highest posterior density (HPD) intervals. If the posterior distribution is skewed, then a two-sided interval with equal tail areas will generally contain some parameter values that have lower posterior probability than values outside the interval. An HPD interval does not have this property – it is adjusted so that the probability ordinates at each end of the interval are identical, and hence it is also the narrowest possible interval containing the required probability. Of course if the posterior distribution has more than one mode, then the HPD may be made up of a set of disjoint intervals.

These alternatives are illustrated in Figure 3.4, suggesting that HPD intervals would be preferable – unfortunately they are generally difficult to compute. For normal posterior distributions these intervals require only the use of tables or

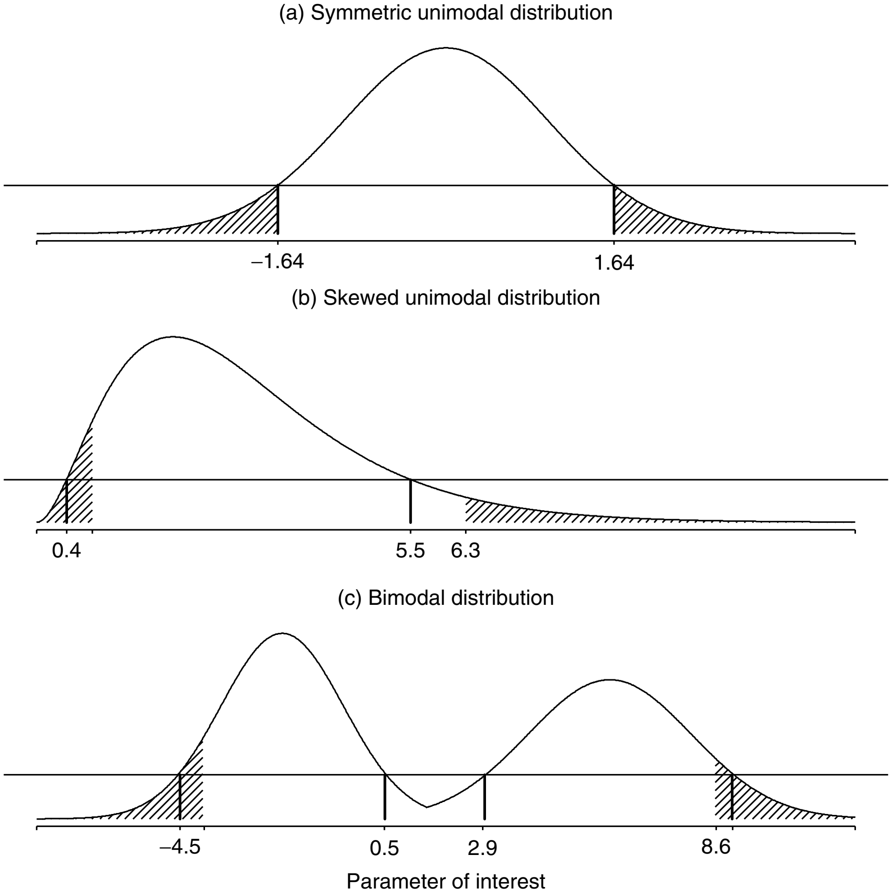


Figure 3.4 (a) shows a symmetric unimodal distribution in which equi-tail-area and HPD intervals coincide at −1.64 to 1.64. (b) is a skewed unimodal distribution in which the equi-tail-area interval is 0.8 to 6.3, whereas the HPD of 0.4 to 5.5 is considerably shorter. (c) shows a bimodal distribution in which the equi-tail-area interval is −3.9 to 8.6, whereas the HPD appropriately consists of two segments.

programs giving tail areas of normal distributions (Sections 2.3 and 3.7). In more complex situation we shall generally be simulating values of θ and one- and two-sided intervals are constructed using the empirical distribution of simulated values (Section 3.19.3). It will not usually be possible to find HPD intervals when using simulation methods.

Traditional confidence intervals and Bayesian credible intervals differ in a number of ways.

1. Most important is their interpretation: we say there is a 95% probability that the true θ lies in a 95% credible interval, whereas this is certainly *not* the

interpretation of a 95% confidence interval. In a long series of 95% confidence intervals, 95% of them should contain the true parameter value – unlike the Bayesian interpretation, we cannot give a probability for whether a *particular* confidence interval contains the true value, it either does or does not and all we have to fall back on is the long-run properties of the procedure. Of course, the direct Bayesian interpretation is often wrongly ascribed to confidence intervals.

2. Credible intervals will generally be narrower due to the additional information provided by the prior: for an analysis assuming the normal distribution they will have width $2 \times 1.96 \times \sigma/\sqrt{n_0 + m}$, compared to $2 \times 1.96 \times \sigma/\sqrt{m}$ for the confidence interval.
3. Some care is required in terminology: while the width of classical confidence intervals is governed by the *standard error* of the estimator, the width of Bayesian credible intervals is dictated by the posterior *standard deviation.*

Interval hypotheses. Suppose a hypothesis of interest comprises an interval $H_0 : \theta_L < \theta < \theta_U$, for some prespecified θ_L, θ_U indicating, for example, a range of clinical equivalence. Then it is straightforward to report the posterior probability $p(H_0|y) = p(\theta_L < \theta < \theta_U|y)$, which may again be obtained using standard formulae or simulation methods.

Example 3.5 *SBP (continued): Interval estimation*

We extend Example 3.4 to encompass testing the hypothesis that the woman has a long-term SBP greater than 135, and the provision of 95% intervals.

The probability of the hypothesis H_0: $\theta_L < \theta < \infty$, $\theta_L = 135$, is

$$p(H_0|y) = p(\theta > \theta_L|y) = 1 - \Phi\left(\frac{\theta_l - \dfrac{n_0\mu + my_m}{n_0 + m}}{\sigma/\sqrt{n_0 + m}}\right)$$

and is shaded in Figure 3.5(a). Figure 3.5(b) displays a 95% posterior interval comprising the posterior mean $\pm 1.96 \times \sigma/\sqrt{n_0 + m}$. Table 3.5 provides the results for both prior and posterior.

We can contrast the Bayesian analysis with the classical conclusions drawn from the likelihood alone. This would comprise a 95% confidence interval $y_m \pm 1.96 \times \sigma/\sqrt{m}$, and a one-sided P-value

$$p(Y < y_m|H_0) = \Phi\left(\frac{y_m - \theta_L}{\sigma/\sqrt{m}}\right);$$

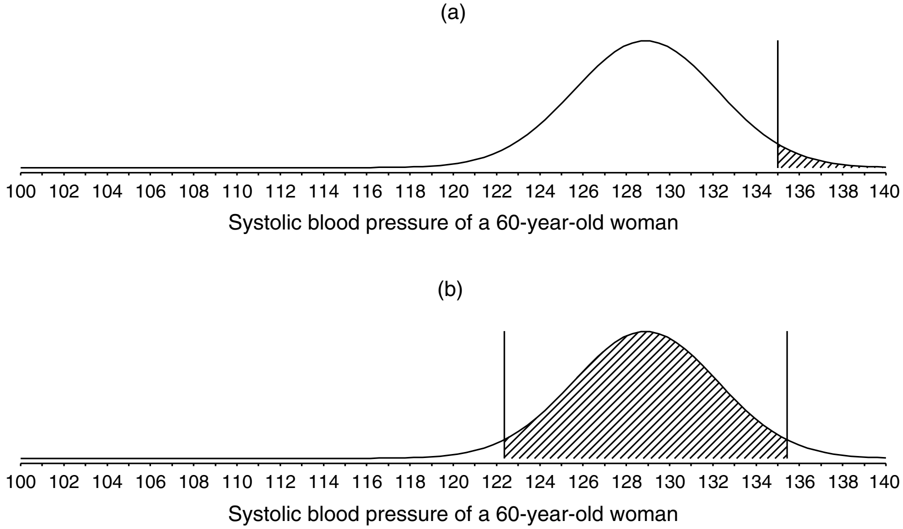


Figure 3.5 Inference from the posterior distribution of the true underlying systolic blood pressure of a 60-year-old woman: (a) shaded area is the probability 0.033 that $\theta > 135$; (b) a two-sided 95% interval (both equi-probability and HPD).

this is numerically identical to the tail area of the posterior with a uniform prior obtained by setting $n_0 = 0$.

We note from Table 3.5 that a traditional one-sided p-value for the hypothesis H_0: $\theta > 135$ is 0.08, while the Bayesian analysis has used the prior opinion to reduce this to 0.03.

Table 3.5 Bayesian and traditional intervals and tests of hypothesis H_0: $\theta > 135$.

	Mean	SD	95% credible interval	$p(H_0 \mid y_m)$
Prior	120.0	10.0	100.4 to 139.6	0.067
Posterior	128.9	3.3	122.4 to 135.4	0.033
	Estimate	SE	95% CI	$p(Y < y_m \mid H_0)$
Classical	130.0	3.5	123.1 to 136.9	0.078

If we were to express the (rather odd) prior belief that all values of θ were equally likely, then $p(\theta)$ would be constant and (3.5) shows that the resulting posterior distribution is simply proportional to the likelihood: (3.14) shows this is equivalent to assuming $n_0 = 0$ in an analysis assuming a normal distribution. In many standard situations a traditional confidence interval is essentially equivalent to a credible interval based on the likelihood alone, and Bayesian and classical results may therefore be equivalent when using a uniform or 'flat' prior. Burton (1994) claims that 'it is already common practice in medical statistics to interpret a frequentist confidence interval as if it did represent a Bayesian posterior probability arising from a calculation invoking a prior density that is uniform on the fundamental scale of analysis'. In our examples we shall present the likelihood and often interpret it as a posterior distribution after having assumed a 'flat' prior: this can be termed a 'standardised likelihood', and some possible problems with this are discussed in Section 5.5.1.

Example 3.6 presents a Bayesian analysis of a published trial: it uses a highly structured format which will be discussed further in Section 3.21. We are aware of the potentially confusing discussion in terms of mortality rates, odds ratios, log(odds ratios) and risk reduction – this multiple terminology is unfortunately inevitable and it is best to confront it early on.

Example 3.6 *GREAT (continued): Bayesian analysis of a trial of early thrombolytic therapy*

Reference: Pocock and Spiegelhalter (1992).

Intervention: Thrombolytic therapy after myocardial infarction, given at home by general practitioners.

Aim of study: To compare anistreplase (a new drug treatment to be given at home as soon as possible after a myocardial infarction) and placebo (conventional treatment).

Study design: Randomised controlled trial.

Outcome measure: Thirty-day mortality rate under each treatment, with the benefit of the new treatment measured by the odds ratio, OR, *i.e.* the ratio of the odds of death following the new treatment to the odds of death on the conventional: OR < 1 therefore favours the new treatment.

Statistical model: Approximate normal likelihood for the logarithm of the odds ratio (Section 2.4).

Prospective Bayesian analysis?: No, it was carried out after the trial reported its results.

Prior distribution: The prior distribution was based on the subjective judgement of a senior cardiologist, informed by empirical evidence derived from one unpublished and two published trials, who expressed belief that 'an expectation of 15–20% reduction in mortality is highly plausible, while the extremes of no benefit and a 40% relative reduction are both unlikely'. This has been translated to a normal distribution on the log(OR) scale, with a prior mean of $\mu_0 = -0.26$ (OR $= 0.78$) and symmetric 95% interval of -0.51 to 0.00 (OR 0.60 to 1.00), giving a standard deviation of 0.13. This prior is shown in Figure 3.6(a).

Loss function or demands: None specified.

Computation/software: Conjugate normal analysis (3.14).

Evidence from study: The 30-day mortality was 23/148 on control and 13/163 on new treatment.

We have already seen in Example 2.5 that the estimated log(OR) is $y_m = -0.74$ (OR $= 0.48$), with estimated standard error 0.36, giving a 95% classical confidence interval for log(OR) from -1.45 to -0.03 (OR from 0.24 to 0.97). The traditional standardised test statistic is therefore $-0.74/0.36 = 2.03$, and the null hypothesis of no effect is therefore rejected with a two-sided P-value of $2\Phi(-2.03) = 0.04$ (GREAT Group, 1992). Figure 3.6(b) shows the likelihood expressing reasonable support for values of θ representing a 40–60% reduction in odds of death. As explained in Example 2.5, it is convenient to express the variance of y_m as σ^2/m, and take $\sigma = 2$ and $m = 30.5$.

Bayesian interpretation: Figure 3.6(c) shows the posterior distribution, obtained by multiplying the prior and likelihood together and then making the total area under the curve equal to one (*i.e.* 'certainty'). The prior distribution has a standard deviation of 0.13, and expressing this as $\sigma/\sqrt{n_0}$ leads to an equivalent number of observations $n_0 = \sigma^2/0.13^2 = 236.7$. Thus the prior can be thought to have around 236.7/30.5 $\approx$ 8 times as much information as the likelihood, showing the strength of the subjective judgement in this example.

The equivalent number of observations in the posterior is then $n_0 + m = 236.7 + 30.5 = 267.2$, with a posterior mean equal to the weighted average $(n_0\mu + my_m)/(n_0 + m) = -0.31$ with standard deviation $\sigma/\sqrt{n_0 + m} = \sigma/\sqrt{267.2} = 0.12$. Thus, the estimated odds ratio is around $e^{-0.31} = 0.73$, or 27% risk reduction (half that observed in the trial). A 95% credible interval can be calculated on the log(OR) scale to be from -0.55 to -0.07, which corresponds to odds ratios from 0.58 to 0.93, or a 95% probability that the true risk reduction lies between 7% and 42%. The posterior probability that the reduction is at least 50% can be calculated by noting this is equivalent to a log(OR) of -0.69, which gives a probability of $\Phi((-0.69 + 0.31)/0.12) = \Phi(-3.11) = 0.001$. We can also calculate the posterior probability that there is any treatment effect as $p(\theta < 0|y_m) = \Phi((0 + 0.31)/0.12) = \Phi(2.54) = 0.995$ and so, adopting the prior provided by the 'expert', we can be 99.5% certain the new treatment is of benefit. Nevertheless, the evidence in the likelihood has been pulled back towards the prior distribution – a formal representation of the belief that the results were 'too good to be true'.

Sensitivity analysis: As an alternative prior formulation, we consider an observer who has no prior bias one way or another, but is more sceptical about large treatment effects than the current expert: this can be

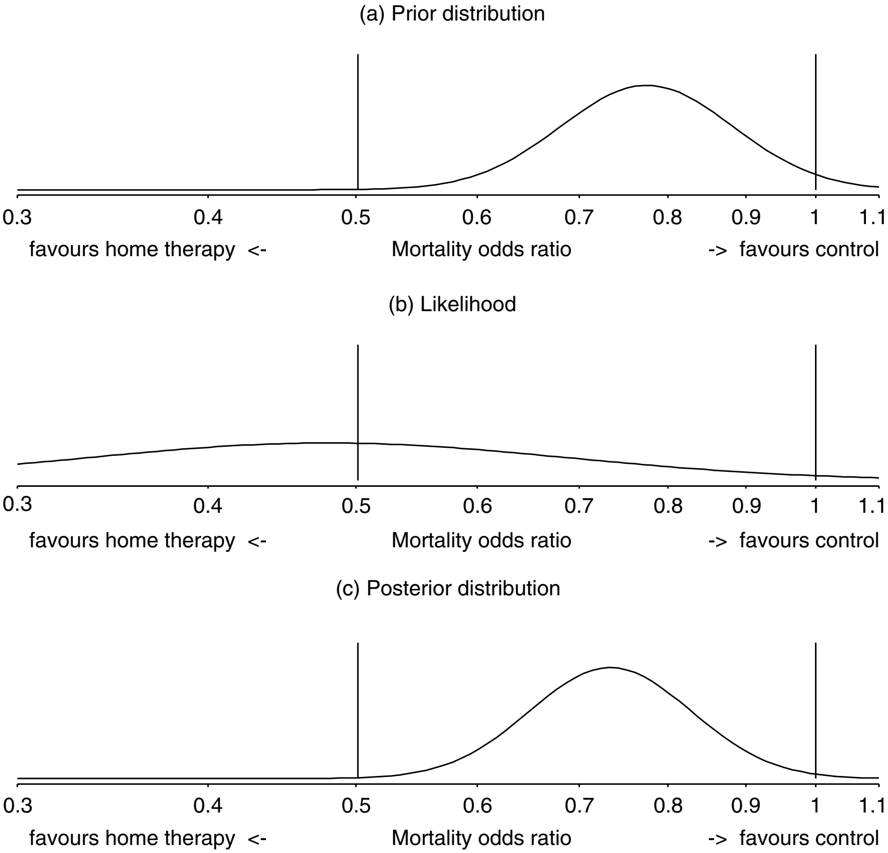


Figure 3.6 Prior, likelihood and posterior distributions arising from GREAT trial of home thrombolysis. These are all normal on the $\theta = \log(\mathrm{OR})$ scale.

represented by a normal prior centred on log(OR) $= 0$ (OR $= 1$) and with a 95% interval that runs from a 50% reduction in odds of death (OR $=$ 0.5, log(OR) $= -0.69$), to a 100% increase (OR $= 2.0$, log(OR) $= 0.69$). On a log(OR) scale, this prior has a 95% interval from -0.69 to 0.69, and so has a standard deviation 0.69/1.96 $= 0.35$ and hence $m = 4/0.35^2 = 32.3$, approximately the same weight of evidence as the likelihood. The prior can therefore be thought of as providing equivalent evidence to that arising from an *imaginary* balanced trial, in which around 16 deaths were observed on each arm. This prior is shown in Figure 3.7, together with the likelihod and posterior distribution, which has mean -0.36 (OR $= 0.70$) and equivalent size $n_0 + m = 62.8$, leading to a standard deviation of 0.25. The probability that there is no benefit from the new treatment is now only $\Phi(-0.36/0.25) = \Phi(-1.42) = 0.08$, shown as the shaded area in Figure 3.7. This analysis suggests that a reasonably sceptical person may therefore not find the GREAT results convincing that there is a benefit: these ideas are formally explored in Section 3.11.

Comments: It is interesting to note that Morrison *et al.* (2000) conducted a meta-analysis of early thrombolytic therapy and estimated OR = 0.83 (95% interval from 0.70 to 0.98), far less impressive than the GREAT results and reasonably in line with the posterior distribution shown in Figure 3.6, which was calculated 8 years before publication of the meta-analysis.

However, this finding should not be over-interpreted and two points should be kept in mind. First, Morrison *et al.* (2000) include some trials that contributed to the prior used by the expert in the above example, and so there is good reason why our posterior (which could be interpreted as a type of subjective meta-analysis) and the formal meta-analysis should correspond. Second, their primary outcome measure is in-hospital mortality, for which GREAT showed a non-significant (but still substantial) benefit of 11/163 vs. 17/148, with an estimated OR of 0.57.

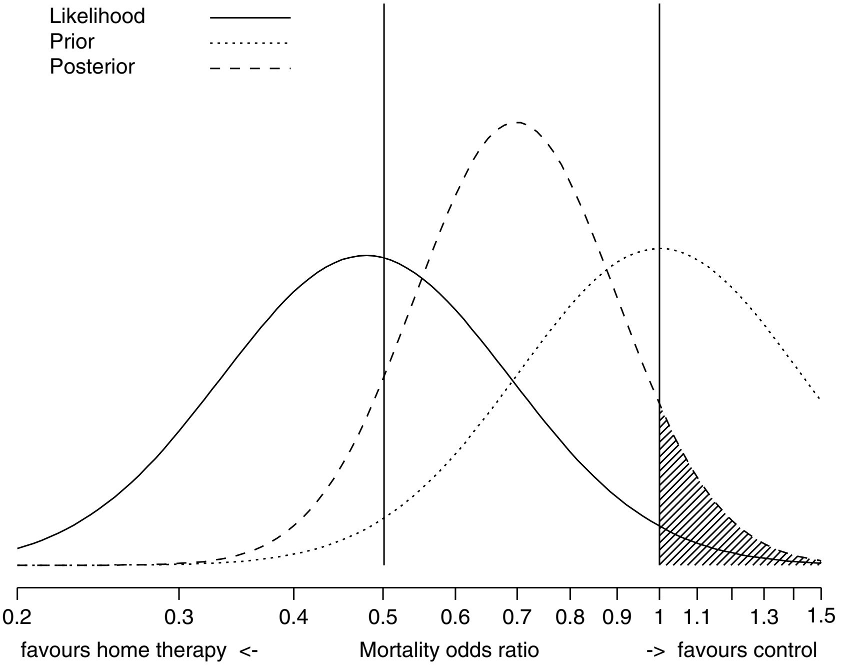


Figure 3.7 A prior distribution that expresses scepticism about large treatment effects would be centred on 0 and have, for example, a 95% interval for OR between 0.5 and 2.0. This is equivalent to a previous study in which 32.3 events occurred, divided equally between the two arms. Adopting this prior and updating it with the GREAT data leads to a posterior distribution as shown, with the shaded area representing a probability of 8% that the treatment is harmful.

3.9 THE PRIOR DISTRIBUTION

Bayesian analysis is driven by the prior distribution, and its source and use present many challenges. These will be covered in detail in Chapter 5, including elicitation from experts, derivation from historical data, the use of 'default' priors to represent archetypal positions of ignorance, scepticism and enthusiasm and, when multiple related studies are being simultaneously analysed, the assumption of a common prior that may be 'estimated'.

It is important to clarify a number of possible misconceptions that may arise. In particular, a prior is:

Not necessarily specified beforehand. Despite the name 'prior' suggesting a temporal relationship, it is quite feasible for a prior distribution to be decided *after* seeing the results of a study, since it is simply intended to summarise reasonable uncertainty given evidence external to the study in question. Cox (1999) states:

> I was surprised to read that priors must be chosen before the data have been seen. Nothing in the formalism demands this. Prior does not refer to time, but to a situation, hypothetical when we have data, where we assess what our evidence would have been if we had had no data. This assessment may rationally be affected by having seen the data, although there are considerable dangers in this, rather similar to those in frequentist theory.

Naturally when making predictions or decisions one's prior distribution needs to be unambiguously specified, although even then it is reasonable to carry out analysis of sensitivity to alternative choices.

Not necessarily unique. There is no such thing as the 'correct' prior. Instead, researchers have suggested using a 'community' of prior distributions expressing a range of reasonable opinions. Thus a Bayesian analysis of evidence is best seen as providing a mapping from specified prior beliefs to appropriate posterior beliefs.

Not necessarily completely specified. When *multiple* related studies are being simultaneously analysed, it may be possible to have unknown parameters in the prior which are then 'estimated' – this is related to the use of hierarchical models (Section 3.17).

Not necessarily important. As the amount of data increases, the prior will, unless it is of a pathological nature, be overwhelmed by the likelihood and will exert negligible influence on the conclusions.

Of course, conclusions strongly based on beliefs that cannot be supported by concrete evidence are unlikely to be widely regarded as convincing, and so it is important to attempt to find consensus on reasonable sources of external evidence. As a true exemplification of the idea that the prior distribution should be under the control of the consumer of the evidence, Lehmann and Goodman (2000) describe ambitious interactive software which allows users to try their own prior distributions.

3.10 HOW TO USE BAYES THEOREM TO INTERPRET TRIAL RESULTS

There have been many connections made between the use of Bayes theorem in diagnostic testing (Example 3.1) and in general clinical research, pointing out that just as the prevalence of the condition (the prior probability) is required for the assessment of a diagnostic test, so the prior distribution on θ should supplement the usual information (P-values and confidence intervals) which summarises the likelihood. We need only think of the huge number of clinical trials that are carried out and the few clearly beneficial interventions found, to realise that the 'prevalence' of truly effective treatments is low. We should thus be cautious about accepting extreme results, such as observed in the GREAT trial, at face value; indeed, it has been suggested that a Bayesian approach provides 'a yardstick against which a surprising finding may be measured' (Grieve, 1994b). Example 3.7 illustrates this need for caution.

Example 3.7 *False positives: 'The epidemiology of clinical trials'*

Simon (1994b) considers the following (somewhat simplified) situation. Suppose 200 trials are performed, but only 10% are of truly effective treatments. Assume each trial is carried out with Type I error α of 5% (the chance of claiming an ineffective treatment is effective) and Type II error β of 20% (the chance of claiming an effective treatment is ineffective) – these are typical values adopted in practice. Table 3.6 displays the expected outcomes: of the 180 trials of truly ineffective treatments, 9 (5%) are expected to give a 'significant' result; similarly, of 20 trials of effective treatments, 4 (20%) are expected to be negative.

Table 3.6 shows that $9/25 = 36\%$ of trials with significant results are in fact of totally ineffective treatments: in diagnostic testing terms, the 'predictive

Table 3.6 The expected results when carrying out 200 clinical trials with $\alpha = 5\%, \beta = 20\%$, and of which only 10% of treatments are truly effective.

		Treatment		
		Truly ineffective	Truly effective	
Trial conclusion	Not significant	171	4	175
	Significant	9	16	25
		180	20	200

value positive' is only 64%. In terms of the odds formulation of Bayes theorem (3.2), when a 'significant result' is observed,

$$\frac{p(H_0|\text{'significant result'})}{p(H_1|\text{'significant result'})} = \frac{p(\text{'significant result'}|H_0)}{p(\text{'significant result'}|H_1)} \times \frac{p(H_0)}{p(H_1)}$$
$$= \frac{p(\text{Type I error})}{1 - p(\text{Type II error})} \times \frac{p(H_0)}{p(H_1)}.$$

Hence the prior odds 0.90/0.10 on the treatment being ineffective (H_0) are multiplied by the likelihood ratio $\alpha/(1 - \beta) = 0.05/0.80 = 1/16$ to give the posterior odds 9/16, corresponding to a probability of 9/25.

Qualitatively, this says that if truly effective treatments are relatively rare, then a 'statistically significant' result stands a good chance of being a false positive.

The analysis in Example 3.7 simplistically divides trial results into 'significant' or 'non-significant', the Bayes factor (likelihood ratio) for the null hypothesis is $\alpha/(1 - \beta)$: this might typically be $0.05/0.80 = 1/16$, categorised as 'strong' evidence against H_0 by Jeffreys (see Table 3.2). However, in Section 4.4.2 we describe how the relationship between Bayes factors and traditional hypothesis tests depends crucially on whether one knows the precise *P*-value or simply whether a result is 'significant'. We note that Lee and Zelen (2000) suggest selecting α so that the posterior probability of an effective treatment, having observed a significant result, is sufficiently high, say above 0.9. This is criticised by Simon (2000) and Bryant and Day (2000) as being based solely on whether the trial is 'significant' or not, rather than the actual observed data.

3.11 THE 'CREDIBILITY' OF SIGNIFICANT TRIAL RESULTS*

We have already seen in Example 3.6 how a 'sceptical' prior can be centred on 'no treatment difference' ($\theta = 0$) to represent doubts about large treatment effects. It is natural to extend this approach to ask how sceptical we would have to be *not* to find an apparently positive treatment effect convincing (Matthews, 2001). Specifically, suppose we have observed data y which is apparently 'significant' in the conventional sense, in that the classical 95% interval for θ based on a normal likelihood lies wholly above or below 0. In addition, suppose our prior mean is 0, reflecting initial scepticism about treatment differences, with the variance of the prior expressing the degree of scepticism with which we view extreme treatment effects, either positive or negative. Matthews (2001) derives an expression for the critical prior distribution which would just lead to the corresponding posterior 95% interval including 0.

Suppose we observe $y_m < 0$. For a normal likelihood and prior with mean 0, (3.14) shows that

$$\theta \sim \mathrm{N}\left[\frac{my_m}{n_0+m}, \frac{\sigma^2}{n_0+m}\right],$$

which means that the upper point u_m of the 95% posterior interval is

$$u_m = \frac{my_m}{n_0+m} + 1.96\frac{\sigma}{\sqrt{n_0+m}}.$$

The 95% interval will therefore overlap 0 if $u_m > 0$. Simple rearrangement shows this will happen provided

$$n_0 > \left(\frac{my_m}{1.96\sigma}\right)^2 - m = \frac{m^2}{1.96^2\sigma^2}\left(y_m^2 - \frac{1.96^2\sigma^2}{m}\right), \tag{3.16}$$

which provides a simple formula for determining the effective number of events in the sceptical prior that would just lead to a 95% posterior interval including 0.

Matthews (2001) shows that we can work directly in terms of the lower and upper points of a 95% interval based on the data alone, denoted l_D and u_D. Thus $l_D, u_D = y_m \pm 1.96\sigma/\sqrt{m}$. It follows that $(u_D - l_D)^2 = 4 \times 1.96^2\sigma^2/m$, and $u_Dl_D = y_m^2 - 1.96^2\sigma^2/m$. Then from (3.16) the critical value of n_0 occurs when the lower point of the 95% prior interval, $l_0 = -1.96\sigma/\sqrt{n_0}$, obeys

$$l_0 = \frac{-1.96\sigma}{\sqrt{n_0}} = -\frac{(u_D - l_D)^2}{4\sqrt{u_Dl_D}}.$$

Often we will be working, say, on a log(odds ratio) scale: if we let $l_0 = \log(L_0), l_D = \log(L_D), u_D = \log(U_D)$ then the corresponding expression is

$$L_0 = \exp\left(\frac{-\log^2(U_D/L_D)}{4\sqrt{\log(U_D)\log(L_D)}}\right). \tag{3.17}$$

L_0 is the critical value for the lower end of a 95% sceptical interval, such that the resulting posterior distribution has a 95% interval that just includes 1. Thus if one's prior belief lies wholly within $(L_0, 1/L_0)$ then one will not be convinced by the evidence, and Matthews suggests a significant trial result is not 'credible' unless prior experience indicates that odds ratios lying outside this critical prior interval are plausible. Figure 3.8 describes how this can be applied to assessment of 'significant' odds ratios.

Applying Figure 3.8 to the GREAT study, for which $L_D = 0.24, U_D = 0.97$, gives $L_0 = 0.10$. Hence, unless odds ratios more extreme than 0.1 can be considered as plausible, the results of the GREAT study should be treated with caution. Since such values do not seem plausible, we do not find the GREAT results 'credible'. This is easily seen to be a characteristic of any 'just significant' results such as those observed in the GREAT trial: just a minimal amount of prior scepticism is necessary to make the Bayesian analysis 'non-significant'. Examples of this approach to scepticism are given in Examples 3.8 and 3.13.

Example 3.8 *Credibility: Sumatriptan trial results*

Matthews (2001) considers the results of an early study of subcutaneous sumatriptan for migraine. This was a small study in which 79% of patients receiving sumatriptan reported an improvement compared to 25% with a placebo, with an estimated odds ratio in favour of sumatriptan of 11.4 and a wide 95% interval of 6.0 to 21.5: the likelihood is shown in Figure 3.9, and we note that odds ratios greater than 1 favour the new

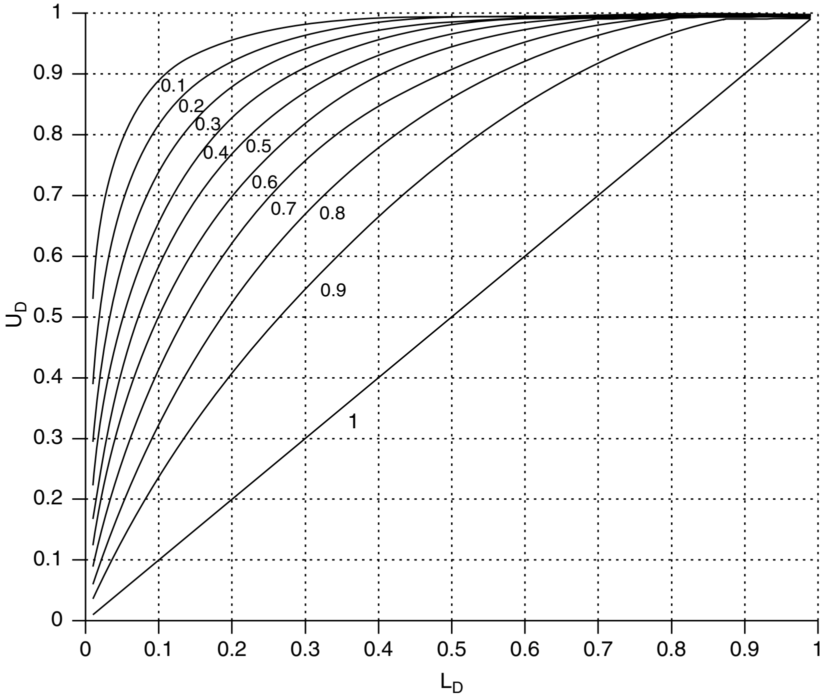


Figure 3.8 Assessment of 'credibility' of findings. Suppose one had observed a classical 95% interval (L_D, U_D) for an odds ratio. Then the value given in the graph is L_0, which is the lower end of a 95% prior interval centred on 1 expressing scepticism about large differences. L_0 is the critical value such that the resulting posterior distribution has a 95% interval that just includes 1, and hence does not produce 'convincing' evidence. Thus, unless values for the odds ratio more extreme than L_0 are judged plausible based on evidence external to the study, then the 'significant' conclusions should not be considered convincing.

treatment since in this application the events are ‘positive’. It is reasonable to ask whether such extreme results are really ‘too good to be true’. To use Figure 3.8 or (3.17) we first need to invert to odds ratios in favour of placebo, *i.e.* ORs less than 1: this leads to an estimated odds ratio of 0.088 with an interval (L_D, U_D) of (0.05, 0.17). Examination of Figure 3.8 reveals an approximate L_0 of 0.8: substitution in (3.17) gives an exact value of $L_0 = 0.84$. Transforming back to the original definition of the odds ratio gives a critical prior interval of $(1/L_0, L_0) = (0.84, 1/0.84) = (0.84, 1.19)$. Figure 3.9 shows this critical prior and the resulting posterior distribution whose 95% interval just includes OR = 1.

If 95% of our prior belief lies within this critical interval, then the posterior 95% interval would not exclude OR = 1 and we would not find the data convincing. However, it would seem unreasonable in this context to rule out on prior grounds advantages of greater than 19%, and hence we reject this critical prior interval as being unreasonably sceptical, and accept the results as ‘credible’.

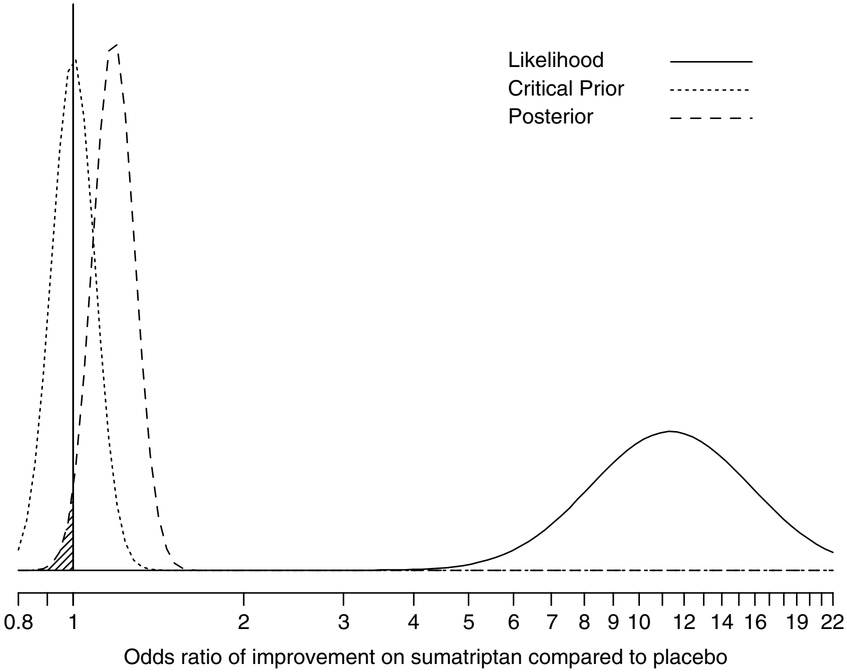


Figure 3.9 Sumatriptan example: the critical sceptical prior distribution (dotted) is centred on OR = 1 and is sufficiently sceptical to make the resulting posterior distribution have a 95% interval that just includes 1, *i.e.* the shaded area is 0.025. However, this degree of prior scepticism seems unreasonably extreme, and hence we might judge that the clinical trial findings are ‘credible’.

3.12 SEQUENTIAL USE OF BAYES THEOREM*

Suppose we observe data in two or more segments, say y_m followed by y_n. Then after the first segment is observed our posterior distribution is given by (3.5):

$$p(\theta|y_m) \propto p(y_m|\theta)\ p(\theta). \tag{3.18}$$

This posterior becomes the prior distribution for the next use of Bayes theorem, so after the next segment y_n is observed, the posterior conditioning on all the data, *i.e.* $p(\theta|y_n, y_m)$, obeys

$$p(\theta|y_n, y_m) \propto p(y_n|\theta, y_m)\ p(\theta|y_m). \tag{3.19}$$

Combination of the two expressions (3.18) and (3.19) yields

$$p(\theta|y_n, y_m) \propto p(y_n|\theta, y_m)\ p(y_m|\theta)\ p(\theta);$$

this can also be derived by considering a single use of Bayes theorem with data y_n, y_m, but factorising the joint likelihood as $p(y_n, y_m|\theta) = p(y_n|\theta, y_m)p(y_m|\theta)$. In most situations the first term in (3.19) will not depend on y_m (*i.e.* Y_n is conditionally independent of Y_m given θ (Section 2.2.3)) and so $p(\theta|y_m)$ simply becomes the prior for a standard Bayesian update using the likelihood $p(y_n|\theta)$.

Example 3.9 *GREAT (continued): Sequential use of Bayes theorem*

Suppose the GREAT trial in Example 3.6 had a first analysis around half way through the trial with the results shown in Table 3.7(b). The estimated log(OR), its standard error and the effective number of events assuming $\sigma = 2$ are calculated as in Example 2.5, and are presented in Table 3.7 with the prior mean and effective number of events in the prior derived in Example 3.6. Bayes theorem assuming normal likelihoods leads to the posterior distribution shown in Table 3.7(c): as shown in (3.14), the effective number of events has been added to $236.7 + 18.1 = 254.8$, and the posterior mean is the weighted average of the prior and likelihood estimates $(236.7 \times -0.255) + (18.1 \times -0.654)/254.8 = -0.283$. The posterior standard deviation is obtained as $\sigma/\sqrt{254.8} = 0.125$.

The second half of the study then provided the data shown in Table 3.7(d), which made up the final totals of 23/144 under control and 13/163 under the new treatment. The sequential use of Bayes theorem means that the posterior following the first part of the study simply becomes the prior for the second, and the final posterior distribution arises in the same manner as described above.

Table 3.7 Possible results were the GREAT trial to have been analysed midway: the 'final' posterior is based on using the posterior from the first part of the trial as the prior for the second part, while the 'combined' posterior is based on pooling all the data into the likelihood. The results only differ through inadequacy of the normal approximation.

Stage	Control deaths/ cases	New treatment deaths/ cases	Estimated log(OR)	Effective no. events	Estimated SE
(a) Prior			−0.255	236.7	0.130
(b) Data – first half	13/74	8/82	−0.654	18.1	0.471
(c) Interim Posterior			−0.283	254.8	0.125
(d) Data – second half	10/74	5/81	−0.817	13.1	0.552
(e) 'Final' posterior			−0.309	267.9	0.122
(f) Combined data	23/144	13/163	−0.736	30.5	0.362
(g) 'Combined' posterior			−0.309	267.2	0.122

We note that the results obtained by carrying out the analysis in two stages (effective number of events 267.9) do not precisely match those obtained by using the total data shown in Table 3.7(g) (effective number of events 267.2). This is due to the quality of the normal approximation to the likelihood when such small numbers of events are observed.

3.13 PREDICTIONS

3.13.1 Predictions in the Bayesian framework

Making predictions is one of the fundamental objectives of statistical modelling, and a Bayesian approach can make this task reasonably straightforward. Suppose we wish to predict some future observations x on the basis of currently observed data y. Then the distribution we require is $p(x|y)$, and (2.8) shows we can extend the conversation to include unknown parameters θ by

$$p(x|y) = \int p(x|y,\theta)\ p(\theta|y)\ d\theta.$$

Now our current uncertainty concerning θ is expressed by the posterior distribution $p(\theta|y)$, and in many circumstances it will be reasonable to assume that x

and y are conditionally independent given θ, and hence $p(x|y,\theta) = p(x|\theta)$. The predictive distribution thus becomes

$$p(x|y) = \int p(x|\theta)\, p(\theta|y)\, d\theta,$$

the sampling distribution of x averaged over the current beliefs regarding the unknown θ. Provided we can do this integration, prediction becomes straightforward.

Such predictive distributions are useful in many contexts: Berry and Stangl (1996a) describe their use in design and power calculations, model checking, and in deciding whether to conduct a future trial, while Grieve (1988) provides examples in bioequivalence, trial monitoring and toxicology. Applications of predictions considered in this book include power calculations (Section 6.5), sequential analysis (Section 6.6.3), health policy-making (Section 9.8.4), and payback from research (Section 9.10).

3.13.2 Predictions for binary data*

Suppose θ is the true response rate for a set of Bernoulli trials, and that the current posterior distribution for θ has mean μ (note this might be a prior or posterior distribution, depending on whether data has yet been observed). We intend to observe a further n trials, and wish to predict Y_n, the number of successes. Then from the iterated expectation (2.13) given in Section 2.2.2 we know that

$$E(Y_n) = E_\theta[E(Y_n|\theta)] = E_\theta[n\theta] = n\mu, \tag{3.20}$$

which means, in particular, that the probability that the next observation ($n = 1$) is a success is equal to μ, the current posterior mean of θ. For example, after the single observation in Example 3.2, the probability that the next case shows a response is the current posterior mean of θ, *i.e.*

$$\begin{aligned} P(Y_1 = 1) = \mathrm{E}(Y_1) &= \sum_j \theta_j\; p(\theta_j|data) \\ &= (0.2 \times 0.1) + (0.4 \times 0.2) + (0.6 \times 0.3) + (0.8 \times 0.4) = 0.6. \end{aligned}$$

If our current distribution for θ is a conjugate Beta$[a, b]$, we can write down an expression for the exact predictive distribution for Y_n: this is known as the beta-binomial distribution and is given by

$$p(y_n) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \binom{n}{y_n} \frac{\Gamma(a+y_n)\,\Gamma(b+n-y_n)}{\Gamma(a+b+n)}. \tag{3.21}$$

From (3.20) and the fact that $\mathrm{E}(\theta) = a/(a+b)$, we immediately see that the mean of this distribution is

$$\mathrm{E}(Y_n) = n\frac{a}{a+b}.$$

We can also obtain the variance by using the expression for the iterated variance (2.14) given in Section 2.2.2, to give

$$\mathrm{V}(Y_n) = \frac{nab}{(a+b)^2}\frac{a+b+n}{(a+b+1)}. \tag{3.22}$$

We note two special cases of the beta-binomial distribution (3.21). First, when $a = b = 1$, the current posterior distribution is uniform and the predictive distribution for the number of successes in the next n trials is uniform over $0, 1, \ldots, n$. Second, when predicting the next single observation ($n = 1$), (3.21) simplifies to a Bernoulli distribution with mean $a/(a+b)$.

Suppose, then, we start with a uniform prior for θ and then observe m trials, all of which turn out to be positive, so that our posterior distribution is now Beta$[m+1, 1]$ (Section 3.6.2). Then the probability that the event will occur at the next trial is $m/(m+1)$. This is known as 'Laplace's law of succession', and it means that even if an event has happened in every case so far (*e.g.* the sun rising every morning), we can still never be completely certain that it will happen at the next opportunity (that the sun will rise tomorrow).

Example 3.10 shows that the beta-binomial distribution can be used in designing experiments allowing for uncertainty in the true response rate.

Example 3.10 *Drug (continued): Making predictions for binary data*

In Example 3.3 we assumed an initial prior distribution for a drug's response rate that could be approximated by a Beta[9.2,13.8], and then observed 15/20 successes, leading to a posterior Beta[24.2,18.8] shown in Figure 3.10(a). The mean of this posterior distribution is 0.56, and hence from (3.20) this is the predictive probability that the next case responds successfully.

If we plan to treat 40 additional cases, then the predictive distribution of the total number of successes out of 40 is a beta-binomial distribution (3.21) which is shown in Figure 3.10(b), and has mean 22.5 and standard deviation 4.3.

Suppose we would consider continuing a development programme if the drug managed to achieve at least a further 25 successes out of these 40 future trials. The chance of achieving this number can be obtained by summing the probabilities in the right-hand tail of Figure 3.10(b), and comes to 0.329. In Example 3.15 we shall contrast this exact analysis with an approximation using simulation methods.

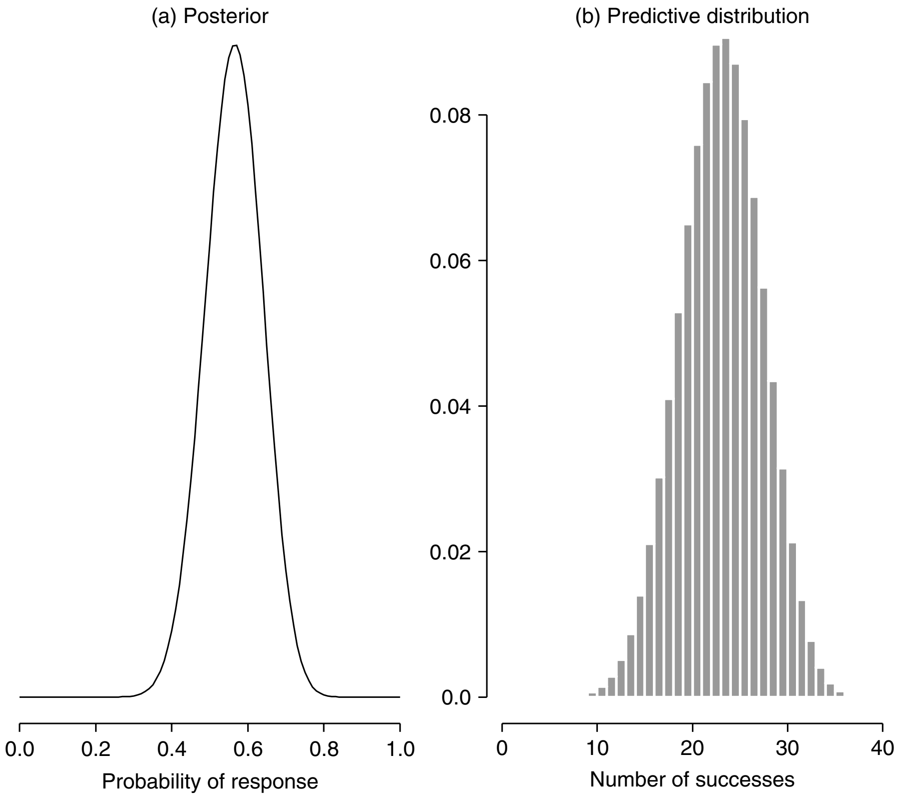


Figure 3.10 (a) is the beta posterior distribution after having observed 15 successes in 20 trials, (b) is the predictive beta-binomial distribution of the number of successes *Y* in the next 40 trials.

3.13.3 Predictions for normal data

Predictions are particularly easy when we are able to assume normal distributions. For example, suppose we assume a normal sampling distribution $Y_n \sim \mathrm{N}[\theta, \sigma^2/n]$ for some future data Y_n, and a prior distribution $\theta \sim \mathrm{N}[\mu, \sigma^2/n_0]$. We wish to make predictions concerning future values of Y_n, taking into account our uncertainty about its mean θ. We may write $Y_n = (Y_n - \theta) + \theta$, and so can consider Y_n as being the sum of two independent quantities: $Y_n - \theta \sim \mathrm{N}[0, \sigma^2/n]$, and $\theta \sim \mathrm{N}[\mu, \sigma^2/n_0]$. Now in Section 2.3 we observed that the sum of two independent normal quantities was normal with the sum of the means and the variances, and hence Y_n will therefore have a predictive distribution

$$Y_n \sim \mathrm{N}\left[\mu, \sigma^2\left(\frac{1}{n} + \frac{1}{n_0}\right)\right]. \tag{3.23}$$

We could also derive (3.23) using the expressions for the iterated expectation (2.13) and variance (2.14) given in Section 2.2.2. Specifically,

$$\begin{aligned} \mathrm{E}(Y_n) &= \mathrm{E}_\theta[\mathrm{E}(Y_n|\theta)] = \mathrm{E}_\theta[\theta] = \mu, \\ \mathrm{V}(Y_n) &= \mathrm{V}_\theta[\mathrm{E}(Y_n|\theta)] + \mathrm{E}_\theta[\mathrm{V}(Y_n|\theta)] = \mathrm{V}_\theta[\theta] + \mathrm{E}_\theta[\sigma^2/n] = \sigma^2(1/n_0 + 1/n). \end{aligned}$$

Thus, when making predictions, we add variances and so *increase* our uncertainty. This is in direct contrast to combining sources of evidence using Bayes theorem, when we add precisions and *decrease* our uncertainty (Section 3.7). The use of this expression for comparison of prior distributions with data is described in Section 5.8, and for sample-size determination in Section 6.5.

Now suppose we had already observed data y_m and hence our distribution is $\theta \sim \mathrm{N}[(n_0\mu + my_m)/(n_0 + m), \sigma^2/(n_0 + m)]$. Then

$$Y_n|y_m \sim \mathrm{N}\left[\frac{n_0\mu + my_m}{n_0 + m}, \sigma^2\left(\frac{1}{n_0 + m} + \frac{1}{n}\right)\right]. \tag{3.24}$$

The use of this expression is illustrated in Example 3.11, and we shall see in Section 6.6.3 how to adapt these methods to predict the chance of a 'significant result' in a clinical trial setting.

Example 3.11 *GREAT (continued): Predictions of continuing the trial*

Suppose we were considering extending the GREAT trial to include a further 100 patients on each arm. What would we predict the observed OR in those future patients to be, with and without using the pre-trial prior information? It is important to remember that the precision with which the OR can be estimated does not depend on the actual number randomised (100 in each arm), but on the number of events (deaths) observed.

We assume the observed log(OR) in those future patients to be $Y_n \sim \mathrm{N}[\theta,\ \sigma^2/n]$, where the future number of events is n and $\sigma = 2$: with 100 patients in each arm we can expect $n \approx 20$ events, given the current mortality rate of around 10%. From Example 3.6, the current posterior distribution is $\theta \sim \mathrm{N}[-0.31,\ \sigma^2/(n_0 + m)]$ where $n_0 + m = 267.2$. Hence from (3.24) the predictive distribution of log(OR) has mean -0.31 and variance $\sigma^2(1/267.2 + 1/20.0) = \sigma^2/18.6 = 0.21 = 0.46^2$. This is shown in Figure 3.11: the great uncertainty in future observations is apparent.

Using the data from the trial alone is equivalent to setting $n_0 = 0$ and using a 'flat' prior, and hence the current posterior distribution is based on the likelihood alone, $\theta \sim \mathrm{N}[-0.74,\ \sigma^2/m]$, where $m = 30.5$. Hence, ignoring the pre-trial prior based on the expert opinion, the predictive distribution of log(OR) has mean -0.74 and variance $\sigma^2(1/30.5 + 1/20.0) = \sigma^2/12.1$

$= 0.33 = 0.58^2$. Figure 3.11 shows that this predictive distribution is considerably flatter than when the prior is included.

We can use the predictive distributions to calculate the chance of any outcome of interest, say observing an OR of less than 0.50 in the future component of the trial. Using the fairly sceptical prior information, this probability is $p(Y_n < \log(0.50)|y_m) = \Phi((-0.69 + 0.31)/0.46) = \Phi(-0.83) = 0.21$, whereas if the prior distribution is ignored this rises to $\Phi((-0.69 + 0.74)/0.58) = \Phi(0.08) = 0.53$. So our prior opinion leads us to doubt that the current benefit will be observed in future patients if the trial is extended.

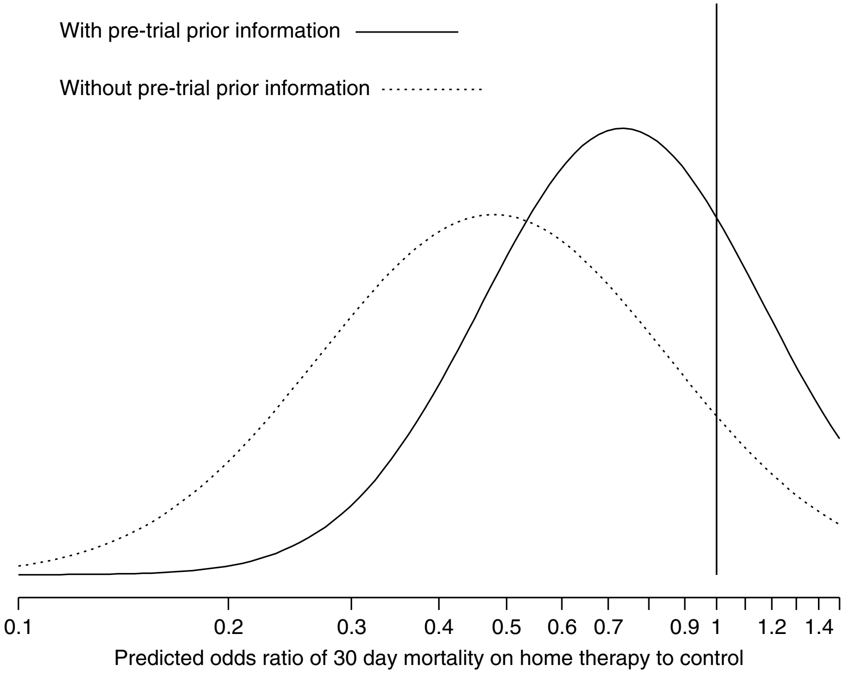


Figure 3.11 Predictive distributions for observed OR in a future 100 patients randomised to each arm in the GREAT trial, assuming around 20 events will be observed: with and without pre-trial prior information.

3.14 DECISION-MAKING

The appropriate role for formal decision theory in health-care evaluation is the subject of a long and continuing debate but is not the primary emphasis of this book. This section presents the basic ideas of which some are developed in later chapters, but for a full discussion we refer to classic texts such as DeGroot

(1970) and Lindley (1975), while Parmigiani (2002) provides a detailed exposition in a medical context.

Suppose we wish to make one of a set of decisions, and that we are willing to assess some value $u(d,\theta)$, known as a *utility*, of the consequences of taking each decision d when θ is the true unknown 'state of nature'. If we have observed some data y and our current probability distribution for θ is $p(\theta|y)$, then our expected utility of taking decision d is denoted

$$\mathrm{E}(d) = \int u\ (d,\theta)\ p(\theta|y)\ d\theta,$$

where the integral is replaced by a sum if θ is discrete. The theory of optimal decision-making says we should choose the decision d^{opt} that maximises $\mathrm{E}(d)$.

For example, suppose our unknown 'state of nature' comprises two hypotheses H_0 and H_1 with current posterior probabilities $p(H_0|y)$ and $p(H_1|y)$ respectively, and assume we face two possible decisions d_0 and d_1: we would choose d_0 if we believed H_0 to be true and d_1 if we believed H_1. Let $u(d_0,H_0)$ be the utility of taking decision d_0 when H_0 is true, and similarly define the other utilities. Then the theory of maximising expected utility states that we should take decision d_0 if $\mathrm{E}(d_0) > \mathrm{E}(d_1)$, which will occur if

$$u(d_0,H_0)p(H_0|y) + u(d_0,H_1)p(H_1|y) > u(d_1,H_0)p(H_0|y) + u(d_1,H_1)p(H_1|y),$$

which can be rearranged to give

$$\frac{p(H_0|y)}{p(H_1|y)} > \frac{u(d_1,H_1) - u(d_0,H_1)}{u(d_0,H_0) - u(d_1,H_0)}. \qquad (3.25)$$

This inequality has an intuitive explanation. The numerator on the right-hand side is $u(d_1,H_1) - u(d_0,H_1)$, the additional utility involved in taking the correct decision when H_1 turns out to be the correct hypothesis – it could also be considered as the potential *regret*, in that it is the potential loss in utility when we erroneously decide on H_0 instead of H_1. The denominator similarly acts as the potential regret when H_0 is true. Hence (3.25) says we should only take decision d_0 if the posterior odds in favour of H_0 are sufficient to outweigh any extra potential regret associated with incorrectly rejecting H_1.

An alternative framework for using the principle of maximising expected utility occurs when our utility depends on future events, and our choice of action changes the probability of those events occurring. Suppose decision d_i can be taken at cost c_i, and leads to a probability p_i of an adverse event $Y = 0$ or 1 occurring with utility U_Y. Then the expected utility of taking decision i is

$$\mathrm{E}(d_i) = p_iU_1 + (1 - p_i)U_0 - c_i,$$

and so, for example, d_0 will be preferred to d_1 if

$$p_0U_1 + (1 - p_0)U_0 - c_0 > p_1U_1 + (1 - p_1)U_0 - c_1.$$

Rearranging terms leads to a preference for d_0 if

$$p_1 - p_0 > \frac{c_0 - c_1}{U_0 - U_1} \tag{3.26}$$

where the denominator $U_0 - U_1$ is positive since the event is considered undesirable. This is clearly obeyed if d_0 both costs less ($c_0 < c_1$) and reduces the risk of Y occurring ($p_0 < p_1$), since the right-hand side of (3.26) is negative and the left-hand side is positive. However, if d_0 costs more than d_1, then the right-hand side of (3.26) is positive, and d_0 will only be preferred if it reduces the risk by a sufficient quantity. We note that the decision depends on the risk difference $p_1 - p_0$, rather than a relative measure such as the odds ratio, and this led Ashby and Smith (2000) to show that (3.26) can be expressed as

$$\text{NNT} = \frac{1}{p_1 - p_0} < \frac{U_0 - U_1}{c_0 - c_1}. \tag{3.27}$$

NNT denotes the 'number needed to treat' in order to prevent one adverse event (the expected number of events prevented when treating N individuals according to d_0 instead of d_1 is $N(p_1 - p_0)$, and hence one expects to prevent one event when treating $N = 1/(p_1 - p_0)$). So, if we are willing to assess the necessary costs and utilities to place in (3.27), we obtain a threshold for adopting a new treatment based on the NNT, without regard to any measure of 'significance'. Example 3.12 provides a somewhat stylised example.

Example 3.12 *Neural tube defects: Making personal decisions about preventative treatment*

Ashby and Smith (2000) consider a somewhat simplified example, but one that nevertheless illustrates the power (and the difficulties) of carrying out a formal decision analysis with utilities.

They consider a couple wishing to try and become pregnant but faced with the decision whether to take folic acid supplements to reduce the risk of a neural tube defect (NTD), such as spina bifida or anencephaly. Let d_0, d_1 denote respectively the decisions to take and not to take supplementation, with respective costs c_0, c_1, and let p_0, p_1 be the probabilities of a foetus having an NTD following each of the two decisions. Finally, let U_0, U_1 be the utilities of having a child without and with an NTD, respectively. The problem is structured as a decision tree in Figure 3.12.

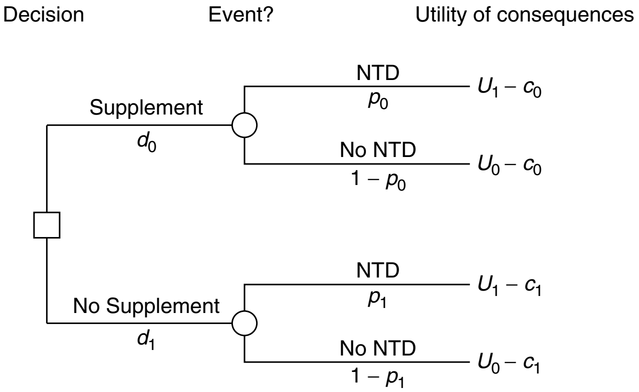


Figure 3.12 Decision tree for folic acid supplementation decision: the square node represents a decision, circular nodes represent chance events, and values at the end of branches represent utilities.

Inequality (3.26) can be rearranged to show that the couple should choose supplementation (d_1) if

$$U_0 - U_1 > \frac{c_0 - c_1}{p_1 - p_0}, \tag{3.28}$$

and the issue becomes one of assigning reasonable values to these quantities. Estimates of p_0 and p_1 may be obtained from randomised trial and epidemiological evidence. Ashby and Smith (2000) provide the results of the sole available clinical trial of folic acid supplementation (carried out on couples who had already had a previous pregnancy resulting in an NTD): 21/602 randomised to placebo had pregnancies with an NTD, compared with 6/593 with supplementation. This corresponds to estimates of $p_0 = 0.010$, $p_1 = 0.035$, NNT $= 1/(p_1 - p_0) = 40.4$ and OR $= 0.30$. Suppose such a couple are deciding whether to take supplementation at a cost of $c_0 - c_1 = £c$; then (3.28) shows they should take the supplementation if the 'disutility' $U_0 - U_1$ of an NTD is greater than around $40c$. c may be costed in money terms if the couple will have to pay for a course of tablets, but Ashby and Smith (2000) suggest this may only be around £10, leading to a threshold of around £400. The problem lies in expressing the 'disutility' in £s.

This brings into focus the importance of identifying the appropriate decision-maker whose utilities are to be taken into account. If making public policy decisions regarding supplementation, it is reasonable that prevention of an NTD is worth more than around $40c$, even if the couple decide to terminate the pregnancy. However, from the couple's point of view, it may be best to think in terms of the utility U_0 of a 'healthy baby'. If this is of the order of £1 million, then they should take supplementation if the utility of an NTD is less than £999 600, which would suggest a fairly clear-cut decision. The crucial quantity is seen to be $S = c/U_0$, the cost of supplementation in terms of 'healthy baby' equivalents. Then the decision threshold (3.28) reduces to checking if

$$\frac{U_1}{U_0} < 1 - (S \times \text{NNT}).$$

Thus the previous analysis had $S \approx 0.000\,01$, NNT ≈ 40, and so supplementation is preferred if an NTD is valued at less than 0.9996 of a healthy baby.

Ashby and Smith (2000) also consider a couple with no previous history of an NTD, and they cite an incidence rate of 3.3 per 1000 pregnancies in a non-supplemented population. Taking this value as $p_0 = 0.0010$, and assuming the trial odds ratio applies to this group, leads to an estimate of $p_1 = 0.0033$, so that $p_1 - p_0 = 0.0023$, NNT $= 435$. We should therefore prefer supplementation if $U_1/U_0 < 1 - 0.000\,01 \times 435 \approx 0.996$. This threshold is again likely to be met, and the costs would need to become very substantial before the threshold was crossed into not preferring supplementation.

The use of Bayesian ideas in decision-making is a huge area of research and application, in which attention is more focused on the utility of consequences than the use of Bayesian methods to revise opinions. This activity blends naturally into cost-effectiveness analysis, but nevertheless the subjective interpretation of probability is essential, since the expressions of uncertainty required for a decision analysis can rarely be based purely on empirical data. There is a long history of attempts to apply this theory to medicine, and in particular there is a large literature on decision analysis, whether applied to the individual patient or for policy decisions. The journal *Medical Decision Making* contains an extensive collection of policy analyses based on maximising expected utility, some of which particularly stress the importance of Bayesian considerations. Any discussion of utility assessment must take careful account of the context in which the analysis is taking place, and our discussion is deferred until the chapter on cost-effectiveness and policy (Chapter 9).

There has been a long debate on the use of loss functions (defined as the negative of utility), in parallel to that concerning prior distributions, and some have continually argued that the design, monitoring and analysis of a study must explicitly take into account the consequences of eventual decisions (Berry, 1993). It is important to note that there is also a frequentist theory of decision-making that uses loss functions, but does not average with respect to prior or posterior distributions: the decision-making strategy is generally ‘minimax’ (DeGroot, 1970), where the loss is minimised whatever the true value of the parameter might be. This can be thought of as assuming the most pessimistic prior distribution. Thus ‘ideological’ approaches employing all combinations of the use of prior distributions and/or loss functions are possible: this is further discussed in Section 4.1 and, in the context of clinical trials, in Section 6.2.

It is particularly important to emphasise that the theory of optimal decision-making depends solely on the *expected* benefit, and hence any measures of uncertainty such as intervals or *P*-values are strictly speaking irrelevant, whether conducting clinical trials (Sections 6.2, 6.6.4 and 6.10) or policy-making (Chapter 9). An exception is when a decision can be made to obtain further information, and these ideas can be used for assessing the payback from research (Section 9.10).

3.15 DESIGN

Bayesian design of experiments can be considered as a natural combination of prediction and decision-making, in that the investigator is seeking to choose a design which they predict will achieve the desired goals. Nevertheless Bayesian design tends to be technically and computationally challenging (Chaloner and Verdinelli, 1995) except possibly in situations such as choosing the size of a clinical trial (Section 6.5).

Sequential designs present a particular problem known as ‘backwards induction’, in which one must work backwards from the end of the study, examine all the possible decision points that one might face, and optimise the decision allowing for all the possible circumstances in which one might find oneself. This can be computationally very demanding since one must consider what one would do in *all* possible future eventualities (Section 6.6.4), although approximations can be made such as considering only a single step ahead. A natural application is in dose-finding studies (Section 6.10). Early phases of clinical trials have tended to attract this approach: for example, Brunier and Whitehead (1994) consider the balancing of costs of experimentation and errors in treatment allocation (Section 6.12).

3.16 USE OF HISTORICAL DATA

Historical evidence has traditionally been used to help in the design of experiments and when pooling data in a meta-analysis, but Bayesian reasoning gives it a formal role in many aspects of evaluation. Here we introduce a brief taxonomy of ways in which historical data may be incorporated, which will be further developed in contexts such as the derivation of prior distributions

(Section 5.4), the use of historical controls in clinical trials (Section 6.9), the adjustment of observational studies for potential biases (Section 7.3) and the synthesis of multiple sources (Section 8.4).

We identify six broad relationships that historical data may have with current observations, ranging from being completely irrelevant to being of equal standing, with a number of possible means of 'downweighting' in between. There is an explicit reliance on judgement as to which is most appropriate in any situation.

(a) *Irrelevance*. The historical data provides no relevant information.
(b) *Exchangeable*. Current and past studies are 'similar' in the sense described in Section 3.17, and so their parameters can be considered exchangeable – this is a typical situation in a meta-analysis, and standard hierarchical modelling techniques can be adopted.
(c) *Potential biases*. Past studies are biased, either through lack of quality (internal bias) or because the setting is such that the studies are not precisely measuring the underlying quantity of interest (external bias), or both. The extent of the potential bias may be modelled and the historical results appropriately adjusted.
(d) *Equal but discounted*. Past studies may be assumed to be unbiased, but their precision is decreased in order to 'discount' past data.
(e) *Functional dependence*. The current parameter of interest is a logical function of parameters estimated in historical studies.
(f) *Equal*. Past studies are measuring precisely the parameters of interest and data can be directly pooled – this is equivalent to assuming exchangeability of individuals.

A fuller graphical and technical description of these stages is provided in Section 5.4.

3.17 MULTIPLICITY, EXCHANGEABILITY AND HIERARCHICAL MODELS

Evaluation of health-care interventions rarely concerns a single summary statistic. 'Multiplicity' is everywhere: clinical trials may present issues of 'multiple analyses of accumulating data, analyses of multiple endpoints, multiple subsets of patients, multiple treatment group contrasts and interpreting the results of multiple clinical trials' (Simon, 1994a). Observational data may feature multiple institutions, and meta-analysis involves synthesis of multiple studies.

Suppose we are interested in making inferences on many parameters $\theta_1, \ldots, \theta_K$ measured on K 'units' which may, for example, be true treatment effects in subsets of patients, multiple institutions, or each of a series of trials. We can identify three different assumptions:

1. *Identical parameters*. All the θs are identical, in which case all the data can be pooled and the individual units ignored.
2. *Independent parameters*. All the θs are entirely unrelated, in which case the results from each unit can be analysed independently (e.g. using a fully specified prior distribution within each unit).
3. *Exchangeable parameters*. The θs are assumed to be 'similar' in the following sense. Suppose we were blinded as to which unit was which, and all we had was a label for each, say, A, B, C and so on. Suppose further that our prior opinion about any particular set of θs would not be affected by only knowing the labels rather than the actual identities, in that we have no reason to think specific units are systematically different. A set of random variables $Y_1, \ldots, Y_n$ with this property was termed 'exchangeable' in Section 3.4, equivalent, broadly speaking, to assuming the variables were independently drawn from some parametric distribution with a prior distribution on the parameter. The results of Section 3.4 can be equally applied to exchangeable parameters $\theta_1, \ldots, \theta_K$, and hence under broad conditions an assumption of exchangeable units is mathematically equivalent to assuming the θs are drawn at random from some population distribution, just as in a traditional random-effects model. This can be considered as a common prior for all units, but one with unknown parameters. Note that there does not need to be any actual sampling – perhaps these K units are the only ones that exist – since the probability structure is a consequence of the belief in exchangeability rather than a physical randomisation mechanism. Nor does the distribution have to be something traditional such as a normal (although we shall generally use that assumption in our examples): heavy-tailed or skewed distributions are possible, or 'partitions' that cluster units into groups that are equal or similar. We emphasise that an assumption of exchangeability is a *judgement* based on our knowledge of the context (Section 5.7).

If a prior assumption of exchangeability is considered reasonable, a Bayesian approach to multiplicity is thus to integrate all the units into a single model, in which it is assumed that $\theta_1, \ldots, \theta_K$ are drawn from some common prior distribution whose parameters are unknown: this is known as a hierarchical or multi-level model.

We illustrate these ideas assuming normal distributions. In each unit we shall observe a response Y_k assumed to have a normal likelihood

$$Y_k \sim \mathrm{N}[\theta_k, s_k^2]. \tag{3.29}$$

The three situations outlined above are then treated as follows.

1. *Identical parameters (pooled effect)*. We assume all the θ_k are identical and equal to a common treatment effect μ and, therefore, from (3.29),

$$Y_k \sim \mathrm{N}[\mu, s_k^2].$$

Transforming to the notation $s_k^2 = \sigma^2/n_k$, assuming $\mu \sim \mathrm{N}[0, \sigma^2/n_0]$ and sequential application of Bayes theorem, (3.14) gives a 'pooled' posterior distribution for μ (and hence each of the θ_k) of

$$\mu \sim \mathrm{N}\left[\frac{\sum_k n_k y_k}{n_0 + \sum_k n_k}, \frac{\sigma^2}{n_0 + \sum_k n_k}\right]; \tag{3.30}$$

the posterior mean for μ is equivalent to an overall sample mean, assuming the prior contributes n_0 'imaginary' observations of 0. As $n_0 \to 0$ the prior distribution on μ becomes uniform and the posterior for μ tends to

$$\mu \sim \mathrm{N}\left[\frac{\sum_k n_k y_k}{\sum_k n_k}, \frac{\sigma^2}{\sum_k n_k}\right]. \tag{3.31}$$

Reverting to the original notation $s_k^2 = \sigma^2/n_k$ reveals that

$$\mu \sim \mathrm{N}\left[\frac{\sum_k y_k/s_k^2}{\sum_k 1/s_k^2}, \frac{1}{\sum_k 1/s_k^2}\right], \tag{3.32}$$

where the posterior mean is simply the classical pooled estimate $\hat{\mu}$, which is the average of the individual estimates, each weighted inversely by its variance. A classical test for heterogeneity, *i.e.* whether it is reasonable to assume that all the trials are measuring the same quantity, is provided by

$$Q = \sum_k \frac{n_k}{\sigma^2}(y_k - \hat{\mu})^2, \tag{3.33}$$

or equivalently $Q = \sum_k (y_k - \hat{\mu})^2/s_k^2$, which has a χ^2_{K-1} distribution under the null hypothesis of homogeneity. It is well known that this is not a very powerful test (Whitehead, 2002), and so absence of a significant Q should not necessarily mean that the trial are homogenous.

2. *Independent parameters (fixed effects).* In this case each θ_k is estimated totally without regard for the others: assuming a uniform prior for each θ_k and the likelihood (3.29) gives the posterior distribution

$$\theta_k \sim \mathrm{N}[y_k, s_k^2], \tag{3.34}$$

which is simply the normalised likelihood.

3. *Exchangeable parameters (random effects).* The unit means θ_k are assumed to be exchangeable, and to have a normal distribution

$$\theta_k \sim \mathrm{N}[\mu, \tau^2], \tag{3.35}$$

where μ and τ^2 are 'hyperparameters' for the moment assumed known. After observing y_k, Bayes theorem (3.15) can be rearranged as

$$\theta_k | y_k \sim \mathrm{N}[B_k\mu + (1 - B_k)y_k, \; (1 - B_k)s_k^2], \tag{3.36}$$

where $B_k = s_k^2/(s_k^2 + \tau^2)$ is the weight given to the prior mean. It can be seen that the pooled result (3.32) is a special case of (3.36) when $\tau^2 = 0$, and the independent result (3.34) a special case when $\tau^2 = \infty$.

An exchangeable model therefore leads to the inferences for each unit having *narrower* intervals than if they are assumed independent, but *shrunk* towards the prior mean response. This produces a degree of pooling, in which an individual study's results tend to be 'shrunk' by an amount depending on the variability between studies and the precision of the individual study. B_k controls the 'shrinkage' of the estimate towards μ, and the reduction in the width of the interval for θ_k. If we again use the notation $s_k^2 = \sigma^2/n_k$, $\tau^2 = \sigma^2/n_0$, then $B_k = n_0/(n_0 + n_k)$, clearly revealing how the degree of shrinkage increases with the relative information in the prior distribution compared to the likelihood.

The unknown hyperparameters μ and τ may be estimated directly from the data – this is known as the 'empirical Bayes' approach as it avoids specification of prior distributions for μ and τ. We shall not detail the variety of techniques available as they form part of classical random-effects meta-analysis (Sutton *et al.*, 2000; Whitehead, 2002). However, the simplest is the 'methods-of-moments' estimator (DerSimonian and Laird, 1986)

$$\hat{\tau}^2 = \frac{Q - (K - 1)}{N - \sum_k n_k^2/N}, \tag{3.37}$$

where Q is the test for heterogeneity given in (3.33), and $N = \sum_k n_k$; if $Q < (K - 1)$, then $\hat{\tau}^2$ is set to 0 and complete homogeneity is assumed. This estimator is used in Example 3.13 and in the Exercises, although we describe the use of 'profile-likelihood' in Section 3.18.

Alternatively, μ and τ^2 may be given a prior distribution (known as the 'full Bayes approach') and this is done later in the book, taking particular care in the choice of a prior distribution for the between-unit variation τ (Section 5.7.3). However, the results from either an empirical or full Bayes analysis will often be similar provided each unit is not too small and there are a reasonable number of units.

The use of hierarchical models is later discussed with respect to subset analysis (Section 6.8.1), *N*-of-1 studies (Section 6.11), institutional comparisons (Section 7.4) and meta-analysis (Section 8.2).

Example 3.13 *Magnesium: Meta-analysis using a sceptical prior*

Reference: Higgins and Spiegelhalter (2002).

Intervention: Epidemiology, animal models and biochemical studies suggested intravenous magnesium sulphate may have a protective effect after acute myocardial infarction (AMI), particularly through preventing serious arrhythmias. A series of small randomised trials culminated in a meta-analysis (Teo *et al*., 1991) which showed a highly significant ($P < 0.001$) 55% reduction in odds of death. The authors concluded that 'further large scale trials to confirm (or refute) these findings are desirable', and the LIMIT-2 trial (Woods *et al*., 1992) published results showing a 24% reduction in mortality in over 2000 patients. An editorial in *Circulation* subtitled 'An effective, safe, simple and inexpensive treatment' (Yusuf *et al*., 1993) recommended further trials to obtain 'a more precise estimate of the mortality benefit'. Early results of the massive ISIS-4 trial pointed, however, to a lack of any benefit, and final publication of this trial on over 58 000 patients showed a non-significant adverse mortality effect of magnesium. ISIS-4 found no effect in any subgroups and concluded that 'Overall, there does not now seem to be any good clinical trial evidence for the routine use of magnesium in suspected acute MI' (Collins *et al*., 1995).

Aim of study: To investigate how a Bayesian perspective might have influenced the interpretation of the published evidence on magnesium sulphate in AMI available in 1993. In particular, what degree of 'scepticism' would have been necessary in 1993 not to be convinced by the meta-analysis reported by Yusuf *et al*. (1993)?

Study design: Meta-analysis of randomised trials, allowing for prior distributions that express scepticism about large effects.

Outcome measure: Odds ratio for in-hospital mortality, with odds ratios less than 1 favouring magnesium.

Statistical model: All three approaches to modelling the multiple trials are investigated: (a) a 'pooled' analysis assuming identical underlying effects; (b) a fixed-effects analysis assuming independent, unrelated effects; and (c) a random-effects analysis assuming exchangeable treatment effects. For the last we assume a normal hierarchical model on the log(OR) scale, as given by (3.29) and (3.35). An empirical Bayes analysis is adopted using estimates of the overall mean μ and the between-study standard deviation τ, in order to use the normal posterior analysis given by (3.36).

Prospective analysis?: No.

Prior distribution: For the pooled- and fixed-effects analysis we assume a uniform prior for the unknown effects on the log(OR) scale. The empirical

Bayes analysis does not use any prior distributions on the parameters μ and τ (although the estimate for μ is equivalent to assuming a uniform prior on the log(OR) scale). Sensitivity analysis is conducted using 'sceptical' priors for μ centred on 'no effect'.

Loss function or demands: None.

Computation/software: Conjugate normal analysis.

Evidence from study: Table 3.8 gives the raw data and the estimated log-odds ratios y_k and their standard deviations s_k (Section 2.4.1). The classical test for heterogeneity Q (3.33) is not significant (9.35 on 7 degrees of freedom), and the method-of-moments estimate for τ is 0.29 (3.37). Figure 3.13 shows the profile log(likelihood) which summarises the support from the data for different values of τ, and is derived using the techniques described in Section 3.18.2: superimposed on this plot are the changing parameter estimates for different values of τ. The maximum likelihood estimate is $\hat{\tau} = 0$ although, from the discussion in Section 2.4.1, values for τ with a profile log(likelihood) above $-1.96^2/2 \approx -2$ might be considered as being reasonably supported by the data. $\hat{\tau} = 0$ would not appear to be a robust choice as an estimate since non-zero values of τ, which are well supported by the data, can have a strong influence on the conclusions. We shall assume, for illustration, the method-of-moments estimator $\hat{\tau} = 0.29$.

The results are shown in Figure 3.14. The standard pooled-effect analysis estimates an odds ratio OR = 0.67 (95% interval from 0.52 to 0.86). In the random-effects analysis the estimates of individual trials are 'shrunk' towards the overall mean by a factor given by B_k in Table 3.8, and individual trials have narrower intervals. The estimate of the 'average' effect is less precise, but still is 'significantly' less than 1: estimated odds ratio 0.58 (95% interval from 0.38 to 0.89).

Table 3.8 Summary data for magnesium meta-analysis, showing estimated odds ratios, log(odds ratios) (y_k), standard deviations for log(odds ratios) (s_k), the effective number of events assuming $\sigma = 2$ (n_k), and shrinkage coefficients $B_k = s_k^2/(s_k^2 + \hat{\tau}^2)$. $\hat{\tau}$ is taken to be 0.29.

Trial	Magnesium group		Control group		Estimated log(odds ratio) y_k	Estimated SD s_k	Effective no. events n_k	Shrinkage B_k
	Deaths	Patients	Deaths	Patients				
Morton	1	40	2	36	−0.65	1.06	3.6	0.93
Rasmussen	9	135	23	135	−1.02	0.41	24.3	0.65
Smith	2	200	7	200	−1.12	0.74	7.4	0.86
Abraham	1	48	1	46	−0.04	1.17	2.9	0.94
Feldstedt	10	150	8	148	0.21	0.48	17.6	0.72
Shechter	1	59	9	56	−2.05	0.90	4.9	0.90
Ceremuzynski	1	25	3	23	−1.03	1.02	3.8	0.92
LIMIT-2	90	1159	118	1157	−0.30	0.15	187.0	0.19

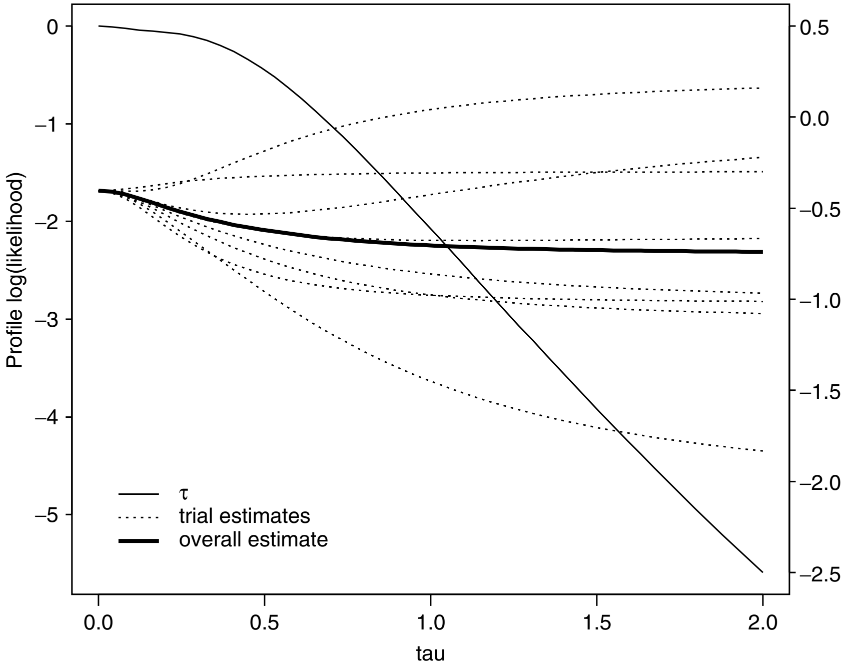


Figure 3.13 Profile log(likelihood) of τ, showing reasonable support for values of τ between 0 and 1. Also shown are individual and overall estimates of treatment effects for different values of τ: although $\tau = 0$ is the maximum likelihood estimate, plausible values of τ have substantial impact on the estimated treatment effects.

Bayesian interpretation: This random-effects analysis is not really a Bayesian technique, as it uses no prior distributions for parameters and conclusions are reported in the traditional way. One could, however, treat this as an approximate Bayesian analysis having assumed exchangeability between treatments and uniform priors on unknown parameters.

Sensitivity analysis: A meta-analysis using uniform prior distributions, whether a pooled- or random-effects analysis, finds a ‘significant’ benefit from magnesium. The apparent conflict between this finding and the results of the ISIS-4 mega-trial have led to a lengthy dispute, briefly summarised in Higgins and Spiegelhalter (2002). We shall return to this issue in Example 8.1, but for the moment we consider the robustness of the meta-analysis results to the choice of prior distribution. In particular, we use the credibility analysis described in Section 3.11 to check whether the findings are robust to a reasonable expression of prior scepticism concerning large benefits. We first consider the pooled analysis. From Figure 3.8, we can see that in order to find unconvincing the

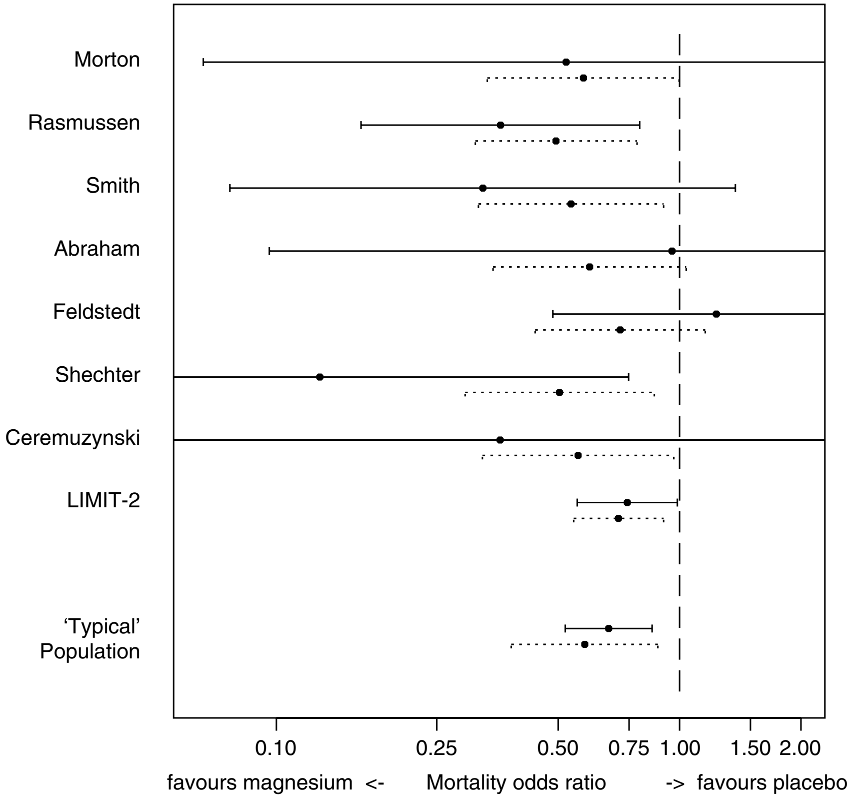


Figure 3.14 Fixed- (solid lines) and random-effects (dashed lines) meta-analysis of magnesium data assuming $\tau = 0.29$, leading to considerable shrinkage of the estimates towards a common value.

pooled analysis (95% interval from 0.52 to 0.86), a sceptical prior with a lower 95% point at around 0.80 would be necessary. Figure 3.15 displays the pooled likelihood, and the 'critical' sceptical prior distribution that leads to a posterior tail area of 0.025 above OR = 1. This prior is N[0, $2^2/421$], and hence is equivalent evidence to a trial in which 421 events have been observed, with exactly the same number in each arm. This seems a particularly extreme form of scepticism in that it essentially rules out all effects greater than around 20% on prior grounds. However, for the random-effects analysis (95% interval from 0.38 to 0.89), the lower end of the sceptical interval would need to be 0.6: the likelihood, 'critical' sceptical prior and posterior are shown in Figure 3.16. It might seem reasonable to find odds ratio below 0.6 extremely surprising, and

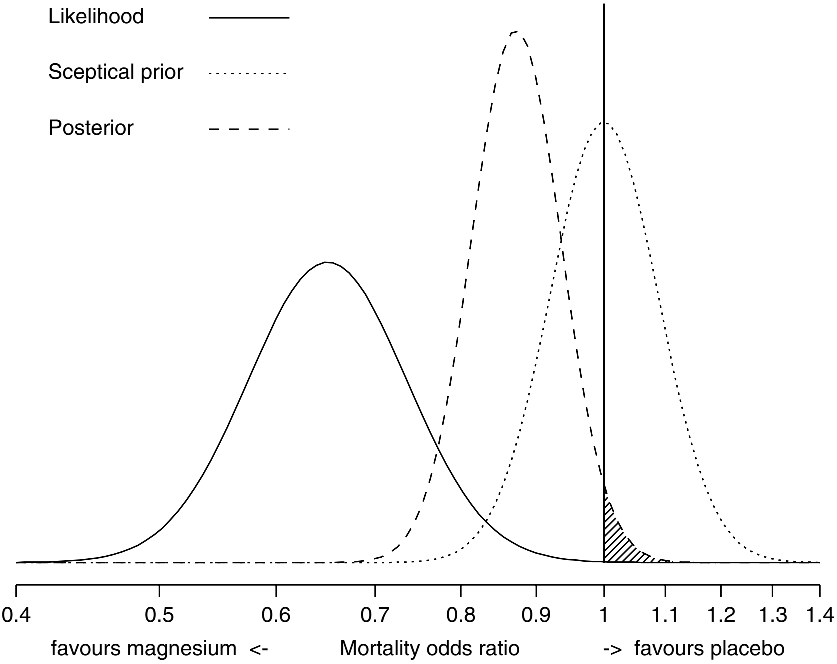


Figure 3.15 Critical sceptical prior for the pooled analysis, just sufficient to make posterior 95% interval include 1. This degree of scepticism seems unreasonably severe, as it equivalent to having already observed 421 events – 210.5 on each treatment.

hence a random-effects analysis and a reasonably sceptical prior render the meta-analysis somewhat unconvincing. This finding is reinforced by the comment by Yusuf (1997) that 'if one assumed that only moderate sized effects were possible, the apparent large effects observed in the meta-analysis of small trials with magnesium ... should perhaps have been tempered by this general judgment. If a result appears too good to be true, it probably is.'

Comments: One vital issue is that the maximum likelihood estimate of τ would lead to assuming a pooled estimate for the odds ratio, whereas there is reasonable evidence for considerable heterogeneity. A simplistic approach in which the maximum likelihood estimate is assumed to be true is therefore likely to substantially overstate the confidence in the conclusions. We note that we might question the exchangeability assumption of a large trial compared with many small ones, and this is further discussed in Higgins and Spiegelhalter (2002).

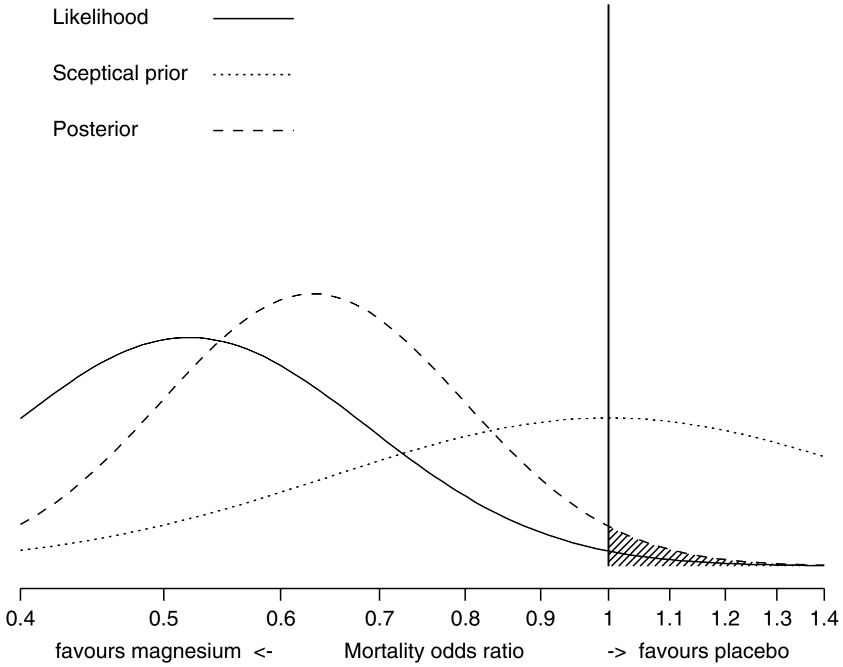


Figure 3.16 Critical sceptical prior for random-effects analysis, just sufficient to make posterior 95% interval include 1. This degree of scepticism appears quite reasonable, corresponding to 58 events (29 in each arm) in a previous 'imaginary trial'.

3.18 DEALING WITH NUISANCE PARAMETERS*

3.18.1 Alternative methods for eliminating nuisance parameters*

In many studies we are focused on inferences on a single unknown quantity θ, such as the average treatment effect in a population of interest. However, there will almost always be additional unknown quantities which influence the data we observe but which are not of primary interest: these are known as 'nuisance' parameters and are a major issue in statistical modeling. Examples include the variance of continuous quantities, coefficients measuring the influence of background risk factors, baseline event rates in control groups, and so on.

Traditional statistical methods are primarily based on analysis of the likelihood for θ, and a number of methods have been developed to eliminate the nuisance parameters from this likelihood. These include the following:

1. Restricting attention to an estimator of θ whose likelihood (at least approximately) does not depend on the nuisance parameters. This technique is used extensively in this book in the form of approximate normal likelihoods for unknown odds ratios, hazard ratios and rate ratios (Section 2.4).
2. Estimating the nuisance parameters and 'plugging in' their maximum likelihood estimates into the likelihood for θ. This ignores the uncertainty concerning the nuisance parameters, and may be inappropriate if the number of nuisance parameters is large. In hierarchical modelling we might use this technique for the hyperparameters of the population distribution, and we saw in Section 3.17 that this is known as the empirical Bayes approach. Example 3.13 showed that conditioning on the maximum likelihood estimate might lead us to ignore an important source of uncertainty.
3. By conditioning on some aspect of the data that is taken to be uninformative about θ, forming a 'conditional likelihood' which depends only on θ.
4. Forming a 'profile likelihood' for θ, obtained by maximising over the nuisance parameters for each value of θ. This was used in Example 3.13 and is illustrated in Section 3.18.2, although here it is not applied to the parameter of primary interest.

Each of these techniques leads to a likelihood that depends only on θ, and which could then be combined with a prior in a Bayesian analysis.

However, a more 'pure' Bayesian approach would be as follows:

1. Place prior distributions over the nuisance parameters.
2. Form a joint posterior distribution over all the unknown quantities in the model.
3. Integrate out the nuisance parameters to obtain the marginal posterior distribution over θ.

This approach features in our examples when we do not assume normal approximations to likelihoods, such as modelling control group risks for binomial data in Examples 8.2 and 9.4, and control group rates for Poisson data in Example 8.3. We also consider full Bayesian modelling of sample variances for normal data in Examples 6.10 and 9.2. In other hierarchical modelling examples we shall generally adopt an approximation at the sampling level, but a full Bayesian analysis of the remaining nuisance parameter: the between-group standard deviation τ.

It is important to emphasise that sensitivity analysis of prior distributions placed on nuisance parameters is important, as apparently innocuous choices may exert unintended influence. For this reason it may be attractive to carry out a hybrid strategy of using traditional methods to eliminate nuisance parameters before carrying out a Bayesian analysis on θ alone, although we might wish to be assured that this was a good approximation to the full Bayesian approach.

3.18.2 Profile likelihood in a hierarchical model*

Consider the hierarchical model described in Section 3.17 and Example 3.13 in which

$$Y_k \sim \mathrm{N}[\theta_k, s_k^2], \qquad \theta_k \sim \mathrm{N}[\mu, \tau^2].$$

The hyperparameters μ and τ^2 will generally be unknown. From (3.24) the predictive distribution of Y_k, having integrated out θ_k, is

$$Y_k \sim \mathrm{N}[\mu, s_k^2 + \tau^2].$$

Let the precision $w_k = 1/(s_k^2 + \tau^2)$ be the 'weight' associated with the kth study. Then the joint log(likelihood) for μ and τ is an arbitrary constant plus

$$L(\mu,\tau) = -\frac{1}{2}\sum_k [(y_k - \mu)^2 w_k - \log w_k]. \tag{3.38}$$

By differentiating (3.38) with respect to μ and setting to 0, we find that, for fixed τ, the conditional maximum likelihood estimator of μ is

$$\hat{\mu}(\tau) = \sum_k y_k w_k / \sum_k w_k, \tag{3.39}$$

with variance $1/\sum_k w_k$ (this is also the posterior mean and variance of μ when assuming a uniform prior distribution for μ). We can therefore substitute $\hat{\mu}(\tau)$ for μ in (3.38) and obtain the profile log(likelihood) for τ as

$$L(\tau) = -\frac{1}{2}\sum_k [(y_k - \hat{\mu}(\tau))^2 w_k - \log w_k]. \tag{3.40}$$

This profile log(likelihood) may be plotted, as in Example 3.13, and maximised numerically to obtain the maximum likelihood estimate $\hat{\tau}$. This can then be substituted in (3.39) to obtain the maximum likelihood estimate of μ.

3.19 COMPUTATIONAL ISSUES

The Bayesian approach applies probability theory to a model derived from substantive knowledge and can, in theory, deal with realistically complex situations – the approach can also be termed 'full probability modelling'. It has to be acknowledged, however, that the computations may be difficult, with the specific problem being to carry out the integrations necessary to obtain the posterior distributions of quantities of interest in situations where non-standard prior distributions are used, or where there are additional 'nuisance parameters' in the model. These problems in integration for many years restricted Bayesian applications to rather simple examples. However, there has recently been enormous progress in methods for Bayesian computation, generally exploiting modern computer power to carry out simulations known as Markov chain Monte Carlo (MCMC) methods (Section 3.19.2).

In this book we shall downplay computational issues and many of our examples can be handled using simple algebra. In practice it is inevitable that MCMC methods will be required for many applications, and our later examples make extensive use of the WinBUGS software (Section 3.19.3).

3.19.1 Monte Carlo methods

Monte Carlo methods are a toolkit of techniques that all have the aim of evaluating integrals or sums by simulation rather than exact or approximate algebraic analysis. The basic idea of replacing algebra by simulation can be illustrated by the simple example given in Example 3.14.

Example 3.14 *Coins: A Monte Carlo approach to estimating tail areas of distributions*

Suppose we want to know the probability of getting 8 or more heads when we toss a fair coin 10 times. An *algebraic* approach would be to use the formula for the binomial distribution given in (2.39) to provide the probability of 8, 9 or 10 heads, which results in

$$\begin{aligned} P(\text{8 or more heads}) &= \binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2+\binom{10}{9}\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right)^1+\binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^0 \\ &= \frac{1}{2^{10}}(45+10+1) \\ &= \frac{56}{1024} \\ &= 0.0547. \end{aligned}$$

An alternative, *physical* approach would be to repeatedly throw a set of 10 coins and count the proportion of throws where there were 8 or more heads. Basic probability theory then says that eventually, after sufficient throws, this proportion will tend to the correct result of 0.0547. This rather exhausting procedure is best imitated by a *simulation* approach in which a computer program generates the throws according to a reliable random mechanism, say by generating a random number U between 0 and 1, and declaring a 'head' if $U \geq 0.5$. The results of 102 such simulated throws of 10 coins are shown in Figure 3.17(a): there were 4, 1 and 0 occurrences of 8, 9 and 10 heads respectively, an overall proportion of $5/102 = 0.0490$, compared to the true probability of 0.0547. Figure 3.17(b) shows the distribution of 10 240 throws, in which there were 428, 87 and 7 occurrences of 8, 9 and 10 heads respectively, instead of the expected counts of 450, 100, and 10. Overall we would therefore estimate the probability of 8 or more heads as $522/10\,240 = 0.0510$. After 10 240 000 simulated throws this empirical proportion is 0.05476, and can be made as close as required to the true value 0.0547 by simply running a longer simulation.

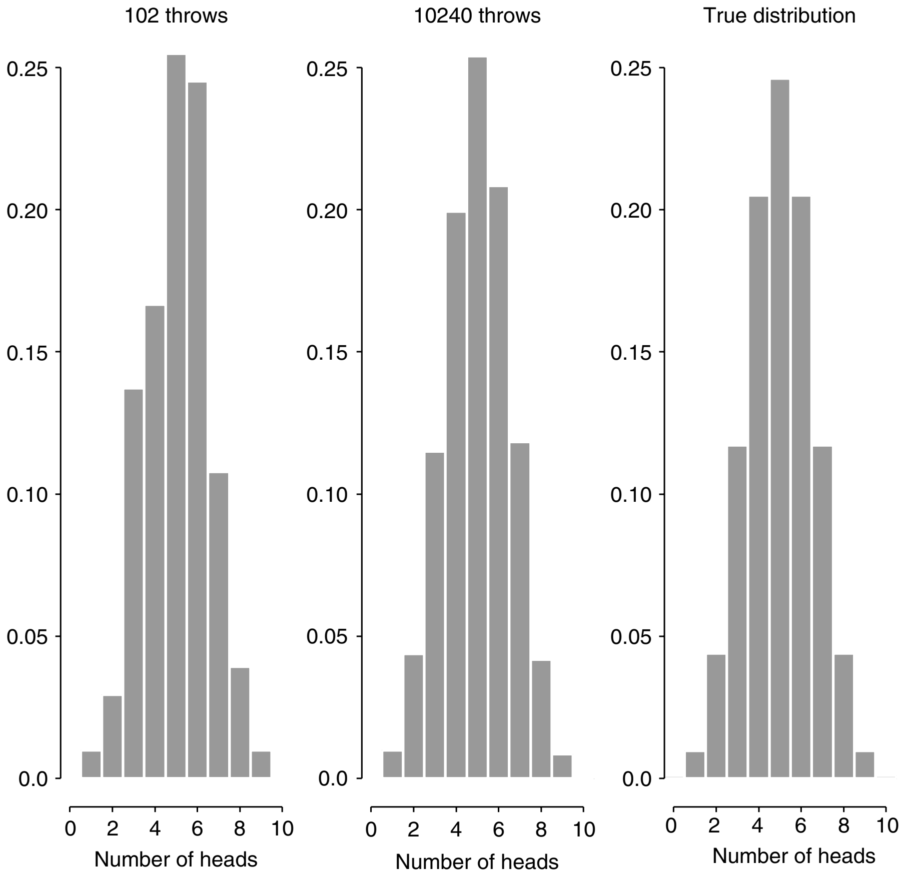


Figure 3.17 (a) Empirical distribution of the number of heads thrown in 102 tosses of 10 balanced coins, where the results of the tosses are obtained by a computer simulation. (b) Empirical distribution after 10 240 throws. (c) True distribution based on the binomial distribution.

The Monte Carlo method described in Example 3.14 is used extensively in risk modelling using software which allows sampling from a wide variety of distributions. The simulated quantities can then be passed into a standard spreadsheet, and the resulting distributions of the outputs of the spreadsheet will reflect the uncertainty about the inputs. This use of Monte Carlo methods can also be termed *probabilistic sensitivity analysis*, and we shall explore this in detail in the context of cost-effectiveness (Section 9.5).

Monte Carlo methods will be useful for Bayesian analysis provided the distribution of concern is a member of a known family – this distribution may be the prior (if no data are available) or current posterior. In conjugate Bayesian analysis it will be possible to derive such a posterior distribution algebraically as in Section 3.6.2 and hence to use Monte Carlo methods to find tail areas (although such tail areas may also be directly obtainable in software), or more usefully to find the distribution of complex functions of one or more unknown quantities as in the probabilistic sensitivity analysis mentioned above. An application of these ideas in power calculations is given in Example 6.5.

3.19.2 Markov chain Monte Carlo methods

Non-conjugate distributions or nuisance parameters (Section 3.18) will generally mean that in more complex Bayesian analysis it will not be possible to derive the posterior distribution in an algebraic form. Fortunately, Markov chain Monte Carlo methods have developed as a remarkably effective means of sampling from the posterior distribution of interest even when the form of that posterior has no known algebraic form. Only a brief overview of these methods can be given here: tutorial introductions are provided by Brooks (1998), Casella and George (1992) and Gilks *et al.* (1996).

The following form the essential components of MCMC methods:

- *Replacing analytic methods by simulation.* Suppose we observe some data y from which we want to make inferences about a parameter θ of interest, but the likelihood $p(y|\theta,\psi)$ also features a set of nuisance parameters (Section 3.18) ψ: for example, θ may be the average treatment effect in a meta-analysis, and ψ may be the control and treatment group response rates in the individual trials. The Bayesian approach is to assess a joint prior distribution $p(\theta,\psi)$, form the joint posterior $p(\theta,\psi|y) \propto p(y|\theta,\psi)p(\theta,\psi)$, and then integrate out the nuisance parameters in order to give the marginal posterior of interest, *i.e.*

$$p(\theta|y) = \int p(\theta, \psi|y)d\psi.$$

In most realistic situations this integral will not be a standard form and some approximation will be necessary. The idea behind MCMC is that we *sample* from the joint posterior $p(\theta,\psi|y)$, and save a large number of plausible values for θ and ψ: we can denote these sampled values as $(\theta^{(1)}, \psi^{(1)})$, $(\theta^{(2)}, \psi^{(2)}), \ldots, (\theta^{(j)}, \psi^{(j)}), \ldots$. Then any inferences we wish to make about θ are derived from the sampled values $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(j)}, \ldots$: for example, we use the sample mean of the $\theta^{(j)}$ as an estimate of the posterior mean $\mathrm{E}(\theta|y)$. We can also create a smoothed histogram of all the sampled $\theta^{(j)}$ in order to estimate the shape of the posterior distribution $p(\theta|y)$. Hence we have replaced analytic integration by empirical summaries of sampled values.

- *Sampling from the posterior distribution.* There is a wealth of theoretical work on ways of sampling from a joint posterior distribution that is known to be proportional to a likelihood × prior, defined as $p(y|\theta,\psi)\ p(\theta,\psi)$, where the latter expression is of known form. These methods focus on producing a *Markov chain*, in which the distribution for the next simulated value $(\theta^{(j+1)}, \psi^{(j+1)})$ depends only on the current $(\theta^{(j)},\psi^{(j)})$. The theory of Markov chains states that, under broad conditions, the samples will eventually converge into an 'equilibrium distribution'. A set of algorithms are available that use the specified form of $p(y|\theta,\psi)p(\theta,\psi)$ to ensure that the equilibrium distribution is exactly the posterior of interest: popular techniques include Gibbs sampling and the Metropolis algorithm, but their details are beyond the scope of this book.
- *Starting the simulation.* The Markov chain must be started somewhere, and *initial values* are selected for the unknown parameters. In theory the choice of initial values will have no influence on the eventual samples from the Markov chain, but in practice convergence will be improved and numerical problems avoided if reasonable initial values can be chosen.
- *Checking convergence.* Checking whether a Markov chain, possibly with very many dimensions, has converged to its equilibrium distribution is not at all straightforward. *Lack* of convergence might be diagnosed simply by observing erratic behaviour of the sampled values, but the mere fact that a chain is moving along a steady trajectory does not necessarily mean that it is sampling from the correct posterior distribution: it might be stuck in a particular area due to the choice of initial values. For this reason it has become generally accepted that it is best to run multiple chains from a diverse set of initial values, and formal diagnostics exist to check whether these chains end up, to expected chance variability, coming from the same equilibrium distribution which is then assumed to be the posterior of interest. This technique is illustrated in Example 3.15, although in the remaining examples of this book we do not go into the details of convergence checking (in fact, our examples are generally well behaved and convergence is not a vital issue).

There are a vast number of published MCMC analyses, many of them using hand-tailored sampling programs. However, the WinBUGS software is widely used in a variety of applications and is essential for many of the examples in this book.

3.19.3 WinBUGS

WinBUGS is a piece of software designed to make MCMC analyses fairly straightforward. Its advantages include a very flexible language for model specification, the capacity to automatically work out appropriate sampling methods, built-in graphics and convergence diagnostics, and a large range of examples and web presence that covers many different subject areas. It has two main disadvantages. The first is its current role as a ‘stand-alone’ program that is not integrated with a traditional statistical package for data manipulation, exploratory analyses and so on (although this is improving to some extent with the ability to call WinBUGS from other statistical packages). Secondly, it assumes that users are skilled at Bayesian analyses and hence can assess the impact of their chosen prior and likelihood, adequately check the fit of their model, check convergence and so on. It is therefore to be used with considerable care. WinBUGS may be obtained from `www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml` (see also Section A.2).

A simple example of the model language was introduced in Example 3.14, which concerned the simulation repeated tosses of 10 ‘balanced coins’. This was carried out in WinBUGS using the program:

```
model{
Y ~ dbin (0.5, 10)
P8 <- step (Y—7.5)
}
```

where Y is binomial with probability 0.5 and sample size 10, and P8 is a step function which will take on the value 1 if Y − 7.5 is non-negative, *i.e.* if Y is 8 or more, 0 if 7 or less. There are only two connectives: The ‘∼’ indicates a distribution, ‘< −’ indicates a logical identity. Running this simulation for 10 240 and 1 024 000 iterations, and then taking the empirical mean of P8, provided the estimated probabilities that *Y* will be 8 or more.

A more complex example is given in Example 3.15, which also illustrates the use of graphs to represent a model, and the use of scripts for running WinBUGS in the background.

Example 3.15 *Drug (continued): Using WinBUGS to implement Markov chain Monte Carlo methods*

In Example 3.10 we used the exact form of the beta-binomial distribution to obtain the predictive distribution of the number of successes in future Bernoulli trials, when the current uncertainty about the probability of success is expressed as a beta distribution. Here we use this example as a demonstration of the ability of the WinBUGS software to both carry out prior-to-posterior analysis and make predictions. In this instance we can compare the results with the exact results derived in Example 3.10; of course, the main use for WinBUGS is in carrying out analyses for which no algebraic solution is possible.

The basic components of the model being considered can be written as

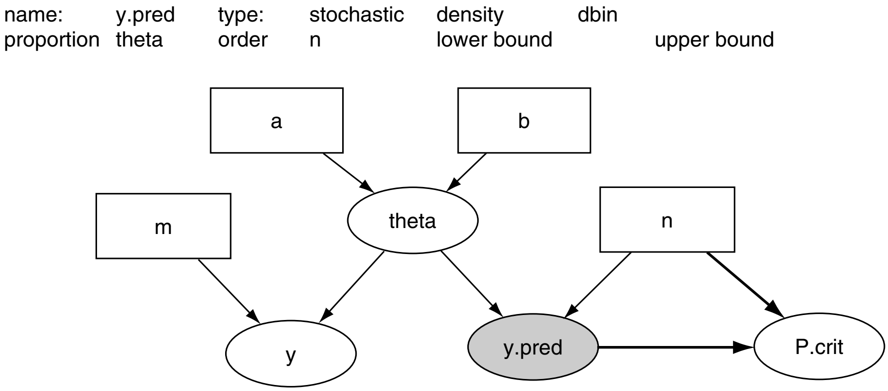
θ	$\sim$	$\text{Beta}[a, b]$	prior distribution
y	$\sim$	$\text{Bin}[\theta, m]$	sampling distribution
y_{pred}	$\sim$	$\text{Bin}[\theta, n]$	predictive distribution
P_{crit}	$=$	$P(y_{\text{pred}} \geq n_{\text{crit}})$	probability of exceeding critical threshold

which is expressed in the WinBUGS language as follows:

```
# WinBUGS analysis of Beta-Binomial 'drug' example
# Model description stored in file 'drug-model.txt'
model{
theta     ~ dbeta(a,b)          # prior distribution
y         ~ dbin(theta,m)       # sampling distribution
y.pred    ~ dbin(theta,n)       # predictive distribution
P.crit    <- step (y.pred-      # =1 if y.pred >= ncrit,
             ncrit+0.5)         # 0 otherwise
}
```

As mentioned in Section 3.19.3, the step function is used here as an indicator as to whether a quantity is greater than or equal to 0, so that the mean of P.crit over a large number of iterations will be the estimate of P_{crit}.

The model is also expressed graphically in Figure 3.18. The representation is described in the figure legend but should be fairly self-explanatory. The important point is that such a directed graph fully describes the joint distribution of all the unknown quantities, and in fact these graphs, known as *Doodles*, can be used by WinBUGS in place of the model syntax above. The part of WinBUGS that deals with the graphs, called DoodleBUGS, can interpret the graphs and either generate WinBUGS code or directly run the



name: y.pred type: stochastic density dbin
proportion theta order n lower bound upper bound

Figure 3.18 Doodle for drug example. The graphical model represents each quantity as a node in directed acyclic graph. Constants are placed in rectangles, random quantities in ovals. Stochastic dependence is represented by a single arrow, and a logical function as a double arrow. The resulting structure is much like a spreadsheet, but allowing uncertainty on the dependencies. WinBUGS allows models to be specified graphically and run directly from the graphical interface.

analysis from the Doodle. Graphical representations can be useful in explaining complex model structures without the distraction of equations; we use them in explaining alternative models for historical data (Section 5.4) and for evidence synthesis (Section 8.4 and Example 8.6).

The relevant values for the model are the parameters of the prior distribution, $a = 9.2$, $b = 13.8$; the number of trials carried out so far, $m = 20$; the number of successes so far, $y = 15$; the future number of trials, $n = 40$; and the critical value of future successes $n_{\text{crit}} = 25$. These values could have been placed in the model description, or alternatively can be written as a list using the format below. This list could be in a separate file or listed after the model description.

```
# data held in file 'data.txt'
# these values could alternatively have been given in model
description
list(
a  = 9.2,      # parameters of prior distribution
b  = 13.8,
y  = 15,       # number of successes
m = 20,        # number of trials
n  = 40,       # future number of trials
ncrit = 25)    # critical value of future successes
```

WinBUGS can automatically generate initial values for the MCMC analysis, but it is better to provide reasonable values in an initial-values list. As mentioned in Section 3.19.2, the best way to check convergence is to carry out multiple runs from widely dispersed starting points and check that, after a suitable 'burn-in', they give statistically indistinguishable chains. This example is simple enough not to require this level of care, but we illustrate the idea by setting up three initial-value files with starting points $\theta = 0.1,\ 0.5,\ 0.9$.

```
# initial values held in file 'drug-in1.txt'
list(theta=0.1)
# initial values held in file 'drug-in2.txt'
list(theta=0.5)
# initial values held in file 'drug-in3.txt'
list(theta=0.9)
```

It is possible to run WinBUGS from a 'point- and-click' interface, but once a program is working it is more convenient to use 'scripts' to carry out a simulation in the background. A script is shown below, checking the syntax of the model, reading in data and multiple initial values, carrying out the simulation and generating the results shown below.

```
# Script for running analysis
display('log')
check('c:/winbugs/drug-model.txt') # check syntax of model
data(' c:/winbugs/drug-dat.txt') # load data file
compile(3) # generate code for 3 simulations
inits(1,'c:/winbugs/drug-in1.txt') # load initial values 1
for theta
inits(2,'c:/winbugs/drug-in2.txt') # load initial values 2
for theta
inits(3,'c:/winbugs' drug-in3.txt') # load initial values 3
for theta
gen.inits() # generate initial value for y.pred
set(theta) # monitor the true response rate
set(y.pred) # monitor the predicted number of successes
set(P.crit) # monitor whether 25 or more successes occur
update(11000) # perform 11000 simulations
```

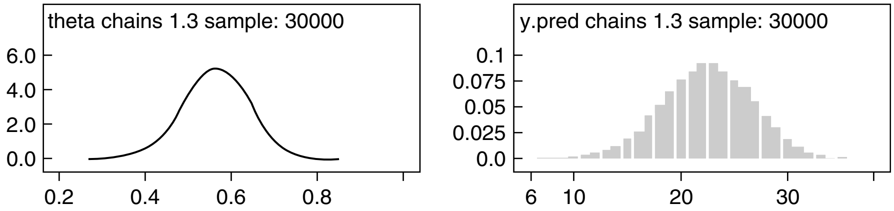


Figure 3.19 Some results based on 30 000 iterations. Convergence is rapidly achieved in such a simple model, and so the burn-in of 1000 iterations was hardly necessary.

```
gr(theta) # Gelman-Rubin diagnostic for convergence
beg(1001) # Discard first 1000 iterations as burn-in
stats(*) # Calculate summary statistics for all monitored
quantities
density(theta) # Plot distribution of theta
density(y.pred) # Plot distribution of y.pred
```

The statistics from the MCMC run are as follows:

```
node    mean    sd      MC error  2.5%    median  97.5%   start sample
P.crit  0.3273 0.4692  0.002631  0.0     0.0     1.0     1001  30000
theta   0.5633 0.07458 4.292E-4  0.4139  0.5647  0.7051  1001  30000
y.pred 22.52   4.278   0.02356  14.0    23.0    31.0     1001  30000
```

The exact answers are available from Example 3.10, and reveal that the posterior distribution has mean 0.563 and standard deviation 0.075, and the beta-binomial predictive distribution has mean 22.51 and standard deviation 4.31. The probability of observing 25 or more successes is 0.329. The MCMC results are within Monte Carlo error of the true values, and can achieve arbitrary accuracy by running the simulation for longer.

The flexibility of WinBUGS allows a variety of modelling issues to be dealt with in a straightforward manner: our examples include inference on complex functions of parameters (Examples 8.4, 8.7 and 9.3), alternative prior distributions (Examples 6.10 and 8.1), inference on ranks (Example 7.2), prediction of effects in new studies (Example 8.1), analysis of sensitivity to alternative likelihood assumptions (Example 8.2), and hierarchical models for both means and variances (Example 6.10).

3.20 SCHOOLS OF BAYESIANS

It is important to emphasise that there is no such thing as a single Bayesian approach, and that many ideological differences exist between researchers. Four broad levels of increasing 'purity' may be identified:

1. The **empirical** Bayes approach (Section 3.17), in which a prior distribution is estimated from multiple experiments. Analyses and reporting are in traditional terms, and justification is through improved sampling properties of procedures.
2. The **reference** Bayes approach, in which a Bayesian interpretation is given to conclusions expressed as posterior distributions, but an attempt is made to use 'objective' or 'reference' prior distributions. There have been a number of attempts to use Bayesian methods but with uniform priors, gaining the intuitive Bayesian interpretation while having essentially the classical results (see Section 5.5; see also Burton *et al.*, 1998; Gurrin *et al.*, 2000). For example, Shakespeare *et al.* (2001) use 'confidence levels' calculated from a normalised likelihood which is essentially a posterior distribution under a uniform prior – this type of activity has been termed an attempt to 'make the Bayesian omelette without breaking the Bayesian eggs'.
3. The **proper** Bayes approach, in which informative prior distributions are based on available evidence, but conclusions are summarised by posterior distributions without explicit incorporation of utility functions. Within this school there may be more or less focus on hypothesis testing using Bayes factors (Section 3.3): Bayes factor analyses essentially entertain the possibility of the precise truth of the null hypothesis (or at least values very close to the null), *i.e. either* θ is extremely close to 0, *or* we have almost no idea of regarding θ. Except in particular circumstances where such dichotomies may be feasible (perhaps in genetics), it might be considered more reasonable to express a 'smooth' sceptical prior: 'in most RCTs, estimation would be more appropriate than testing' (Kass and Greenhouse, 1989).
4. The **decision-theoretic** or 'full' Bayes approach, in which explicit utility functions are used to make decisions based on maximising expected utility. There has been long and vigorous debate on whether or not to incorporate an explicit loss function, and the extent to which a health-care evaluation should lead to an inference about a treatment effect or a decision as to future policy. Important objections to a decision-theoretic approach include the lack of a coherent theory for decision-making on behalf of multiple audiences with different utility functions, the difficulty of obtaining agreed utility values, and the fact that a strict decision-theoretic view would lead to future treatments being recommended on the basis of even marginal expected gains, without any concern as to the level of confidence with which such a recommendation is made (see Section 6.2 and Chapter 9).

Our personal leaning, and the focus in this book, is towards the third, *proper*, school of Bayesianism.

In spite of this apparent divergence in emphasis, the schools are united in their belief in the fundamental importance of three concepts that distinguish Bayesian from conventional methods: *coherence* of probability statements (Section 3.1), *exchangeability* (Section 3.17) and the *likelihood principle* (Section 4.3).

3.21 A BAYESIAN CHECKLIST

Bayesian methods tend to be inherently more complex than classical analyses, and thus there is an additional need for quality assurance. However, there are limited 'guidelines' available for reporting Bayesian analyses. Rudimentary guidance was provided by Lang and Secic (1997), who gave the following instructions:

1. Report the pre-trial probabilities and specify how they were determined.
2. Report the post-trial probabilities and their probability intervals.
3. Interpret the post-trial probabilities.

Similar advice is given in the *Annals of Internal Medicine*'s instructions to authors. The BaSiS (Bayesian Standards in Science) initiative (Section A.2) is seeking to establish guidelines for reporting.

In this section we present a checklist against which published accounts of Bayesian assessments of health-care interventions can be compared. We aim to ensure that an account which adequately contains all the points mentioned here would have the property that the analysis could be replicated by another investigator who has access to the full data. These guidelines should be seen as complementary to the CONSORT (Moher *et al.*, 2001) guidelines, in that they focus on those aspects crucial to an accountable Bayesian analysis, in addition to standard paragraphs concerning the intervention, the design and the results.

Our main examples attempt to use this structure, although it sets a high standard that we admit we do not always reach! In particular, it is often easier to present the evidence at the same time as the statistical model, particularly when there has been some iterative model construction. To avoid tedious repetition, the phrase 'should be clearly and concisely described' should be assumed to apply to each of the components below.

Background

- *The Intervention.* The intervention to be evaluated with regard to the population of interest and so on.
- *Aim of study.* It is important that a clear distinction is made between desired inferences on any quantity or quantities of interest, representing the

parameters to be estimated, and any decisions or recommendations for action to be made subsequent to the inferences. The former will require a prior distribution, while the latter will require explicit or implicit consideration of a loss or utility function.

Methods

- *Study design.* This is a standard requirement, but when synthesising evidence particular attention will be necessary to the similarity of studies in order to justify any assumptions of exchangeability.
- *Outcome measure.* The true underlying parameters of interest.
- *Statistical model.* The probabilistic relationship between the parameter(s) of interest and the observed data, either mathematically, or in such a way as to allow its mathematical form to be unambiguously obtained by a competent reader, including any model selection procedure, whether Bayesian or not.
- *Prospective Bayesian analysis?* It needs to be made clear whether the prior and any loss function were constructed preceding the data collection, and whether analysis was carried out during the study.
- *Prior distribution.* Explicit prior distributions for the parameters of interest should be given. If 'informative', then the derivation of the prior from an elicitation process or empirical evidence should be detailed. If claimed to be 'non-informative', then this claim should be justified. If it is intended to examine the effect of using different priors on the conclusion of the study, this should be stated and the alternative priors explicitly given.
- *Loss function or demands.* An explicit method of deducing scientific consequences is decided prior to the study. This will often be a range of equivalence (a range of values such that if the parameter of interest lies within it, two different technologies may be regarded as being of equal effectiveness), or a loss function whose expected value is to be minimised with respect to the posterior distribution of the parameter of interest. Any elicitation process from experts should be described.
- *Computation/software.* A mathematically competent reader should, if necessary, be able to repeat all the calculations and obtain the required results, and any mathematical software used to obtain the results should be described. If MCMC methods are being used the assumption of convergence should be justified.

Results

- *Evidence from study.* As much information about the observed data – sample sizes, measurements taken – as is compatible with brevity and data confidentiality should be given. It is also essential that the likelihood could be reconstructed, so that subsequent users can establish the contribution from the study to, say, a meta-analysis.

Interpretation

- *Bayesian interpretation*. The posterior distribution should be clearly summarised: in most cases, this should include a presentation of posterior credible intervals and a graphical presentation of the posterior distribution. If either a formal or informal loss function has been described, the results should be expressed in these terms.

  There should be a careful distinction between the report as a current summary for immediate action, in which case a synthesis of all relevant sources of evidence is appropriate, and the report as a contributor of information to a future evidence synthesis.
- *Sensitivity analysis*. The results of any alternative priors and/or expressions of the consequences of decisions.
- *Comments*. These should include an honest appraisal of the strengths and possible weaknesses of the analysis.

3.22 FURTHER READING

Historical references concerning Bayesian methods include Bayes (1763), Holland (1962), Fienberg (1992) and Dempster (1998). For general introductions, see the chapter by Berry and Stangl (1996a) in their textbook (Berry and Stangl, 1996b) which covers a whole range of modelling issues, including elicitation, model choice, computation, prediction and decision-making. Non-technical tutorial articles include Lewis and Wears (1993), Bland and Altman (1998) and Lilford and Braunholtz (1996), while O'Hagan and Luce (2003) provide an excellent primer geared towards cost-effectiveness studies. Other authors emphasise different merits of Bayesian approaches in health-care evaluation: Eddy *et al.* (1990a) concentrate on the ability to deal with varieties of outcomes, designs and sources of bias, Breslow (1990) stresses the flexibility with which multiple similar studies can be handled, Etzioni and Kadane (1995) discuss general applications in the health sciences with an emphasis on decision-making, while Freedman (1996) and Lilford and Braunholtz (1996) concentrate on the ability to combine 'objective' evidence with clinical judgement. Stangl and Berry (1998) provide a recent review of biomedical applications.

There is a huge methodological statistical literature on general Bayesian methods, much of it quite mathematical. Cornfield (1969) provides a theoretical justification of the Bayesian approaches, in terms of ideas such as *coherence*. A rather old article (Edwards *et al.*, 1963) is still one of the best technical introductions to the Bayesian philosophy. Good tutorial introductions are provided by Lindley (1985) and Barnett (1982), while more recent books, roughly in order of increasing technical difficulty, include Berry (1996a), Lee (1997), O'Hagan (1994), Gelman *et al.* (1995), Carlin and Louis (2000), Berger (1985) and Bernardo and Smith (1994).

Recommended references for specific issues include DeGroot (1970) on decision theory, axiomatic approaches and backwards induction, Bernardo and Smith (1994) on exchangeability, and Kass and Raftery (1995) on Bayes factors. On computational issues, Carlin *et al.* (1993) and Etzioni and Kadane (1995) discuss a range of methods which may be used (normal approximations, Laplace approximations and numerical methods including MCMC), Gelman and Rubin (1996) review MCMC methods in biostatistics, and van Houwelingen (1997) provides a commentary on the importance of computational methods in the future of biostatistics.

With regard to hierarchical models Jerome Cornfield (1969, 1976) was an early proponent of the Bayesian approach to multiplicity (Section 6.8.1), while Breslow (1990) gives many examples of problems of multiplicity and reviews the use of empirical Bayes methods for longitudinal data, small-area mapping, estimation of a large number of relative risks in a case–control study, and multiple tumour sites in a toxicology experiment. Louis (1991) reviews the area and provides a detailed case study, while Greenland (2000) provides an excellent justification.

3.23 KEY POINTS

1. Bayesian methods are founded on the explicit use of judgement, formally expressed as prior beliefs and possibly loss functions. The analysis can therefore quite reasonably depend on the context and the audience. However, if the aim is to convince a wide range of opinion, subjective inputs must be strongly argued and be subject to sensitivity analysis.
2. Bayes theorem provides a natural means of revising opinions in the light of new evidence, and the Bayes factor or likelihood ratio provides a scale on which to assess the weight of evidence for or against specific hypotheses.
3. Bayesian methods are best seen as a transformation from initial to final opinion, rather than providing a single ‘correct’ inference.
4. Exchangeability is a vital judgement: exchangeable observations justify the use of parametric models and prior distributions, while exchangeable parameters lead to the use of hierarchical models.
5. Bayesian methods provide a flexible means of making predictions, and this is helped by MCMC methods.
6. Hierarchical models provide a flexible and widely applicable structure when wanting to simultaneously analyse multiple sources of evidence.
7. A decision-theoretic approach may be appropriate where the consequences of a study are considered reasonably predictable, but this is not the emphasis of this book.
8. Normal approximations can be used in many contexts, particularly when deriving likelihoods from standard analyses. This will generally entail transformation between different scales of measurement.

9. Standards for Bayesian reporting have not been established. The most important aspect is to provide details of each of the prior distributions, its justification and its influence assessed through sensitivity analysis.

EXERCISES

3.1. Altman (2001) considers the data in Table 3.9, showing the results of using a scan of the liver to detect abnormalities compared to classification at autopsy, biopsy or surgical inspection in 344 patients.
   (a) Estimate the likelihood ratio for a positive scan.
   (b) For the patients in Table 3.9 the prevalence of an abnormal pathology is 0.75. For this population estimate the posterior probability of an abnormal diagnosis after observing a positive scan result. What is the estimated posterior probability for a population in which the prevalence is 0.25?

3.2. Asked prior to a study of a new chemotherapy, an oncologist said that she would expect 90% of patients to respond, and that she thought it was unlikely to be less than 80%. (a) Use a 'method-of-moments' argument similar to that of Example 3.3 to summarise the oncologist's opinions in terms of a beta distribution, and plot this prior distribution. In 20 patients treated, 14 respond. (b) Plot the likelihood. (c) Update the beta parameters in the light of the data observed.

3.3. Show that if $y_1, \dots, y_n$ are i.i.d. observations from a Poisson distribution with unknown mean θ, and that a gamma prior distribution with parameters α and β is specified for θ, the corresponding posterior distribution is also gamma, i.e. conjugate, with parameters $\alpha + \sum_{i=1}^{n} y_i$ and $\beta + n$.

3.4. Based on national statistics for a large number of similar hospitals, a manager believes that the mean number of patients attending a specialist clinic each week in his hospital should lie between 12 and 20. (a) Taking this range as approximately equivalent to a mean ±2 standard deviations, use a 'method-of-moments' argument similar to that in Example 3.3 to summarise the manager's beliefs using a gamma distribution. The numbers of patients attending a specialist clinic each week for 5 weeks are 11, 15, 18, 13, 19, and are assumed to be independent observations

Table 3.9 Detection of abnormal liver pathology using scan compared to actual classification at autopsy, biopsy or surgical inspection in 344 patients.

	Pathology (Truth)		Total
Liver scan (Test)	Abnormal (+)	Normal (−)	
Abnormal (+)	231	32	263
Normal (−)	27	54	81
Total	258	86	344

from a Poisson distribution. (b) Obtain the posterior distribution for the mean number of patients per week based on the manager's prior beliefs. (c) Plot the prior and posterior densities. If your software permits it, calculate the prior and posterior probabilities that the mean is greater than 18.

3.5. Verify (3.14) algebraically, *i.e.* that a normal prior distribution is conjugate for the unknown mean of a normal likelihood.

3.6. Consider the GREAT trial of home thrombolytic therapy described in Example 3.6. Another cardiologist was more sceptical about the magnitude of benefit and thought that the relative reduction in odds of death was more likely to be around 10–15%, and that the extremes of a 25% relative reduction and a 2.5% increase were unlikely.
   (a) Fit a normal prior distribution for the log(odds ratio) to these opinions.
   (b) Obtain the posterior distribution for this cardiologist and compare it with the posterior distributions in Example 3.6.

3.7. Using the normal approximation to the likelihood derived in Exercise 2.5, assume a sceptical prior distribution, such that an odds ratio of 1 was most likely but with a 95% interval from 0.5 to 2.0. Obtain the posterior estimate for the log(odds ratio), odds ratio and associated 95% intervals.

3.8. Use the normal approximation to the likelihood derived in Exercise 2.8 and assume a sceptical prior distribution equivalent to the evidence in a balanced trial in which 50 events have occurred on each arm. Obtain the corresponding posterior distribution for the log(hazard ratio).

3.9. Using the methods of Section 3.11, consider the results seen in the PROSPER RCT in Exercise 2.8.
   (a) Find the sceptical prior distribution for the log(hazard ratio) with mean 0, such that the resulting posterior 95% interval for the hazard ratio just includes 1.
   (b) Do you think this degree of scepticism is reasonable, and hence are the trial results credible?

3.10. Baum *et al.* (1992) report the results of an RCT to investigate the use of tamoxifen compared to standard care for women treated for breast cancer, evaluated in terms of disease-free survival. In total, 2030 women were randomised and followed up for over 10 years. Overall, there were 484 events in the tamoxifen arm, whilst 419.6 were expected. (a) Assuming balanced randomisation and follow-up, estimate the number of events in the standard-care arm. During the first 5 years of the trial 387 events were observed compared to 320.2 expected, and in the second period of the trial 97 events were observed whilst 99.4 were expected. (b) Assuming a sceptical prior for the log(hazard ratio) centred at zero and with precision equivalent to having observed only 10 events, show that a sequential analysis of the accumulating trial data using the methods of Section 3.12 gives similar results to an analysis using all the trial data.

3.11. In Exercise 2.7 consider another 100 patients randomised between HAI and control.

(a) About how many deaths would we expect to observe?

(b) What would be the predictive distribution for the observed log(hazard ratio) using a sceptical prior distribution, *i.e.* centred at zero and equivalent to having observed 10 deaths?

(c) Repeat (b) for an optimistic prior that represented beliefs that there would be a 10% relative reduction in the risk of death associated with HAI with uncertainty equivalent to having observed 25 deaths.

3.12. Whitehead (2002) considers a meta-analysis of 9 RCTs to evaluate whether taking diuretics during pregnancy reduces the risk of pre-eclampsia and which is summarised in Table 3.10. For each study, (a) estimate the log(odds ratio) and its variance, and (b) obtain an estimate and 95% intervals for the pooled odds ratio. (c) Using the 'method of moments' (3.37), estimate the between-study variance τ^2. Hence obtain the posterior estimates and intervals for (d) the population odds ratio using random effects assuming the between-study variance is known, and (e) the odds ratios for each of the 13 studies assuming a random-effects model.

3.13. Cooper *et al.* (2002) report the results of an economic decision model to assess the cost-effectiveness of using prophylactic antibiotics in women undergoing Caesarean section. Evidence available includes the results of a Cochrane systematic review of 61 RCTs which evaluated the prophylactic use of antibiotics in women undergoing Caesarean section to prevent wound infection, which produces an estimated odds ratio of 0.40, where the baseline probability of wound infection without prophylactic use of antibiotics is estimated to be 0.08. Antibiotic treatment is assumed to cost £10. Women who have a Caesarean section and who do not develop an infection have a mean total cost of £1159 and are

Table 3.10 RCTs evaluating the use of diuretics during pregnancy to reduce risk of pre-eclampsia.

Study	Diuretic		Control	
	Cases	Total	Cases	Total
1	14	131	14	136
2	21	385	17	134
3	14	57	24	48
4	6	38	18	40
5	12	1011	35	760
6	138	1370	175	1336
7	15	506	20	524
8	6	108	2	103
9	65	153	40	102

assumed to have a utility in the subsequent year of 0.95 quality-adjusted life-years (QALYs), while women who have a Caesarean section and who develop an infection have mean total cost of £2320 and utility of 0.80 QALYs: it is assumed there is no difference between the groups after one year.

(a) Structure the decision as in Figure 3.12.

(b) Using the methods of Section 3.14, find the threshold for a policy decision-maker, in £ per QALY, at which the expected utility of using prophylactic antibiotics would exceed that of not using prophylactic antibiotics.

3.14. Use WinBUGS to repeat the analysis of the PROSPER RCT in Exercise 2.9, assuming a uniform prior (on a suitable wide range) for the log(odds ratio), and (a) the approximate normal likelihood, (b) exact binomial likelihoods.

3.15. Use WinBUGS to repeat the analysis in Exercise 3.4 of patients attending a specialist clinic.